SLMs Meet GraphRAG: A Structured Approach to Context-Aware Cybersecurity Hint Generation

Ishan Abraham[1], Jens Mache[1], Taylor Wolff[2], Jack Cook[2], Richard Weiss[2] and Justin Wang[3]
[1]Lewis & Clark College, Portland, Oregon, USA
[2]The Evergreen State College, Olympia, Washington, USA
[3]Northeastern University, Boston, Massachusetts, USA

ishanabraham@lclark.edu
jmache@lclark.edu
taylor.wolff@evergreen.edu
cookjackc@gmail.com
weissr@evergreen.edu
hs.wang@northeastern.edu

Abstract: Generating hints for learners who are engaged in hands-on cybersecurity exercises is the goal of our research. Learners sometimes get stuck or frustrated, they head in the wrong direction or are missing information that is necessary for solving an exercise. While using large language models (LLMs) is an option, LLMs typically require the sharing of student data with third-party AI providers. In order to improve privacy and minimize cost and computational overhead, previous research has explored using locally deployed small language models (SLMs) with retrieval-augmented generation (RAG). However, while RAG has been shown to enhance SLM capabilities without the need to fine tune, it falls short when answering open-ended or multi-step questions that require reasoning across interconnected concepts. This limitation is particularly evident in cybersecurity education, where students often need help understanding how threats, tools, and strategies relate to one another. The cybersecurity hint system EDUHints (Wolff et al, 2025) currently relies on a standard RAG pipeline. In classroom testing, students were unsure whether generated hints meaningfully answered their questions. To address this challenge, we present a custom GraphRAG approach that builds on a proposed cybersecurity education focused ontology and knowledge graph called AISecKG[1]. We extend the ontology to let us incorporate natural language-to-bash command mappings, a valuable feature as students tend to ask questions regarding command-line use. Graph data is extracted using multiple methods and semantically scored to prioritize only the most relevant results. Our pipeline currently employs Microsoft's Phi-3-mini-4k-instruct SLM, integrates LangChain for modular orchestration, and uses Neo4j as the graph database. We survey cybersecurity instructors to rate responses generated by the EDUHints and our GraphRAG system. Results show that hints generated using a GraphRAG are preferred almost three times more by cybersecurity instructors. This suggests that an SLM's educational hint generation abilities can be improved through our GraphRAG architecture.

Keywords: Cybersecurity education, Small language model (SLM), Retrieval-Augmented generation (RAG), Knowledge graph, Human-in-the-Loop

1. Introduction

The rise of LLM chatbots within the educational sphere comes with good reason. These bots offer a great resource for students to receive feedback at times when instructors are not present. However language models may need additional topic and contextual information as they may not have been trained extensively on domain specific data. To mitigate this, researchers and educators have experimented with Retrieval-Augmented Generation[2] (RAG), which allows models to access relevant external information improving answer results. It does this by creating a system where embedded data can be retrieved efficiently and then injected into the model's prompt (Lewis et al, 2020). This serves as a cost effective and scalable way to give the model access to information it was not trained on. RAG as well offers a reliable solution to bolstering the performance of small language models (SLMs) within an educational setting (Liu et al, 2024 and Yu et al, 2025). SLMs also have the benefit of not requiring GPU's and can be run in a fully containerized environment.

With EDURange, a cybersecurity education platform, students learn the fundamentals of command-line cybersecurity while participating in live exercises, letting them apply concepts learned in the classroom to real

[1]Agrawal, G. et al (2023) "AISecKG: Knowledge Graph Dataset for Cybersecurity Education", Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE), https://par.nsf.gov/servlets/purl/10401616

[2]Lewis, P. et al (2020) "Retrieval-augmented generation for knowledge-intensive NLP tasks", Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS), pp 9459–9474, https://dl.acm.org/doi/abs/10.5555/3495724.3496517

applications (Weiss et al, 2017). Previously we introduced EDUHints[3], a human-in-the-loop hint generation system for EDURange, to speed up the instructor feedback process. EDUHints is a system for the instructor that can generate a hint based on a student's most recent bash commands, their questions to instructors, and knowledge check up responses. Instructors also have the ability to manually edit hints before sending them to students ensuring quality and relevancy of hints.

The EDUHint system has thus far been evaluated via three classroom tests, garnering mixed results from students surveyed on the perceived helpfulness and the quality of hints they received (Wolff et al, 2025). These results indicated a clear need for improvements to the system's hint generation process. One such area of improvement that we explore in this work is implementing a more sophisticated retrieval-augmented generation system to improve the capabilities of our SLM.

GraphRAG[4] (Edge et al, 2024) has been shown to be a promising alternative to RAG which allows for our model to understand not just the cybersecurity data, but the relationships between them as well. This tackles a known issue in RAG which is its inability to handle open-ended or multi-step questions (Harjono, 2025). In EDURange, students are required to carry out complex tasks involving knowledge over multiple concepts. Our goal with utilizing a GraphRAG is to accurately convey these relations, providing students more accurate and educational hints.

This paper's main contributions are as follows:

- **Graph-Enhanced Retrieval for SLMs**: We demonstrate that small language models (SLMs) can benefit from a graph-augmented retrieval architecture. By leveraging structured knowledge graphs, our system compensates for the limited reasoning and domain knowledge capabilities typical of SLMs, particularly in cybersecurity education.
- **A Custom GraphRAG Pipeline for Cybersecurity Education**: We design and implement a modular GraphRAG pipeline tailored to EDURange, a real-world cybersecurity learning platform. Our system integrates Phi-3-mini-4k-instruct, Neo4j, and LangChain, and includes custom mechanisms for entity extraction, multi-modal graph querying, vector-based path scoring, and context formatting.

2. Related Work

Recent work highlights that traditional RAG struggles with multi-step reasoning and questions aimed at a broad scope of knowledge. Edge et al. (2024) discussed how RAG only retrieves locally relevant passages without capturing relationships among concepts. Edge also proposes GraphRAG, a graph-based approach that leverages LLM powered community-level summaries and entity-relationship style data. This strategy outperformed standard vector RAG on queries focused on a large corpus of data. Within the education domain, RAG has been shown in tandem with Small Language Models (SLMs) showing promise as a question answering chatbot for introductory computer science (Liu et al, 2024 and Yu et al, 2025) and as a way to improve an SLMs hint generation capabilities in cybersecurity (Wolff et al, 2025). SLMs are used as opposed to LLMs in order to gain control over privacy, cost, as well as local deployment, however their lack of general knowledge leads researchers to find ways to bolster up their domain specific capabilities.

To bring more structured context into retrieval, several groups have focused on domain specific knowledge graphs. Agrawal et al (2023) propose AISecKG, a defined ontology, knowledge graph, and custom entity classification model for the cybersecurity education field. The ontology describes how cybersecurity data can be classified into roles, applications, or concepts as well as further subsections within them. This form of graph centric data enables reasoning entities and their relationships.

Evaluations of RAG-style systems remains an open question. A prominent metric, Ragas[5] (Es et al, 2024) scores how retrieved passages contribute to generated outputs. While this method works well to determine the effectiveness of your knowledge base data, when it comes to educational hints, automated frameworks such as

[3]Wolff, T. et al (2025) "EDUHints: A Human-in-the-Loop Small Language Model System for the Generation of Cybersecurity Hints", Proceedings of the 24th European Conference on Cyber Warfare and Security (ECCWS), https://papers.academic-conferences.org/index.php/eccws/article/view/3659/3360

[4]Edge, D. et al (2024) "From Local to Global: A Graph RAG Approach to Query-Focused Summarization", ArXiv, https://arxiv.org/abs/2404.16130

[5]Es, S. et al (2024) "Ragas: Automated Evaluation of Retrieval Augmented Generation", Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, https://arxiv.org/pdf/2309.15217

this fail to examine the usefulness of a generated response. Similar to past work (Wolff et al, 2025), we survey instructors directly on the preference of hints generated rather than consulting frameworks or relying on LLMs to ensure accurate assessment of the quality of hints.

3. Data/Knowledge Graph

We used an existing cybersecurity education ontology and knowledge graph AISecKG (Agrawal et al, 2023) as the basis for our design. The ontology describes three main categories in which all cybersecurity concepts are categorized in: application, concept, and role. Within each category there are subcategories (12 total) which further define what the data is. The data was created based on 6 upper level cybersecurity lab manuals which outlined many important tools, frameworks, concepts, and how they are used. Data is stored in entity relationship triples making it ideal for a graph database.

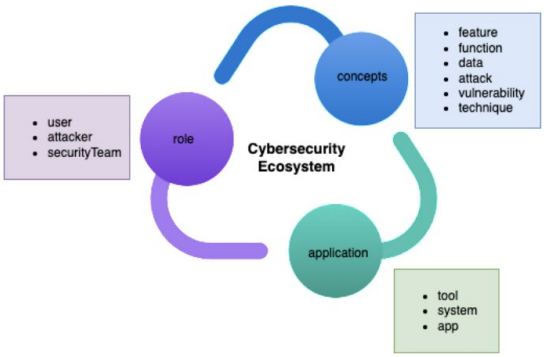


Figure 1: Ontology from AISecKG [Agrawal et al.]

An important piece of data we supplant this ontology with is a newly created natural language to bash commands dataset, NL2SH-ALFA[6] (Westenfelder et al, 2025) . We extend our ontology by adding a category called nl_bash which has two types of data, natural language and bash commands, as well as the relation has_command. We do this to match the triples format to keep data consistent with the rest of the ontology and also to define the connection between the natural language and bash. In total our graph has about 80,100 nodes and 41,200 relationships, with 79,000 of the nodes being either natural language or bash commands. All data has a label, type, category, as well as a vector embedding. We embed all nodes and relationships using the all-MiniLM-L6-v2 embedding model from the sentence transformers library. In the future we hope to improve this by mapping new data to the AISeckKG ontology since the grand majority of our current database is natural language to bash and because we would like more EDURange specific data to be available to our system.

4. Setup/ Implementation

LangChain is an open source framework that allows developers to build applications utilizing all types of language models. LangChain's built-in compatibility with graph databases such as Neo4j, vector databases, and llama-cpp makes it a practical choice for our system. Although LangChain has built in methods for creating a GraphRAG system we found that creating a custom chain was significantly more effective in querying optimal graph data.

[6]https://huggingface.co/datasets/westenfelder/NL2SH-ALFA

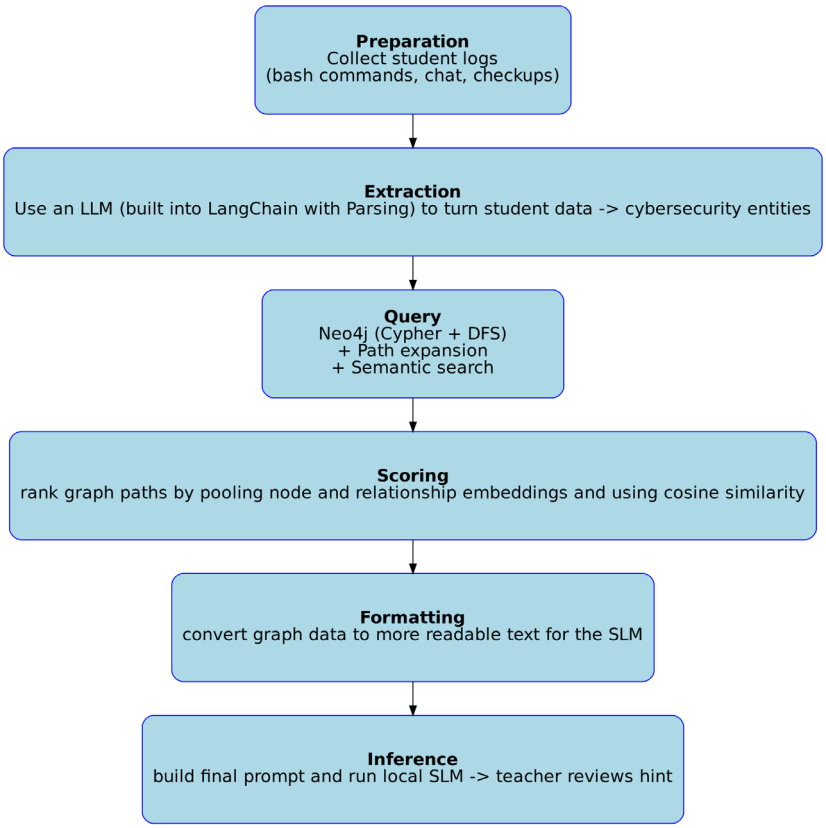


Figure 2: GraphRAG System Design

4.1 Preparation

Unlike question answering systems, the industry standard for educational chatbots, EDUHints and our GraphRAG system work semi-autonomously to collect a student's recent history and use this to generate a hint for the teacher to then review. This human-in-the-loop approach allows the incorporation of artificial intelligence while preserving organic student-instructor interaction. In our preparation step we parse student logs generated by EDURange and store student bash commands, chat questions, and the scenario's check up answers in a dictionary.

4.2 Extraction

Just like other GraphRAG systems, we use an LLM (gpt-3.5-turbo) to process our input and break it into its important cybersecurity relevant entities. We ensure correct formatting by using LangChain's built in PydanticOutputParser and defining a schema for an output, guaranteeing consistency with the LLM output. For example if a student asked the question "What does a firewall do in a network?", the output of our extraction would be ['firewall', 'network'] since these are the relevant cybersecurity terms. We also append the full question as an entity after extraction because much of our natural language to bash command data is in sentence forms.

4.3 Query

We use three separate strategies for obtaining data from our graph while tracking which algorithm was used to query each data point, letting us observe the impacts of each method.

1. **Initial Connection (IC):** We attempt to locate direct paths between two entities in our list via a standard depth first search graph traversal, allowing a max depth of three. IC generally resulted in the fewest amount of queried data for any particular student log. This makes sense as this algorithm requires a connection (of at most three "hops" away) between at least two entities. This strict requirement can often not return any data at all. However, data points found from this method often scored *significantly higher* than the other algorithms as by its nature these data points are most rich with relevant information. We choose this algorithm first to ensure any data captured here is always scored as we limit the total amount of queried data to 20.

2. **Neighborhood Expansion:** Store a two depth "neighbourhood" of each entity. This had shown to consistently output relevant data points, often scoring highest when a learner asks questions about concepts within cybersecurity but Initial Connection returned little to no data. This information provides our SLM with contextual information about the specific concepts a learner is asking about.
3. **Semantic Search:** Since we already have node and relationship embeddings stored in our graph, we embed each entity (which includes the full questions) and use sklearn's NearestNeighbors function to find the top 5 closest matching nodes in our graph to EACH entity we extracted. We then run Neighborhood Expansion on this set of nodes. This approach was shown to excel when students asked questions about bash commands, very likely due to our large amount of natural language to bash command data mappings.

This tri-query system ultimately aims to **minimize retrieval bias**, allowing us to capture the most relevant and useful data that a particular algorithm may miss. However, since the methods run independently, we also must explicitly deduplicate to ensure data is unique.

4.4 Scoring

We limit our querying methods to cap at 20 data points per method, more than enough data for most queries. After combining and deduplicating data we implement a method of ranking these data points so that the most relevant graph data shows up first. We utilize our already existing vector embeddings for all nodes and relationships by taking the mean of all node and relationship embeddings within a path, resulting in a vector which semantically represents a full data point. We can then compare these pooled embeddings to a learner's chat questions via cosine similarity to find the closest matching data points. This showed to be incredibly effective at identifying which graph data was the most important in answering a learner's question. Since our small language model was tested to be volatile with longer context windows we only choose the top 4 highest scoring data points. One notable edge case is the situation where a student has not asked any questions yet chooses to generate a hint. In this case, we skip scoring altogether as in order for our scoring mechanism to work, we require at least one question in order to have a comparison for our queried graph data.

4.5 Formatting

We organize graph data into an ordered list with the most relevant data point at the top. We as well format our data to be significantly more context window friendly while maintaining all information. This is done by converting Neo4j's dictionary data points into a more human readable string format.

4.6 Inference

Once we have our formatted data we can build our final prompt which includes student bash commands, chat history, check up answers, as well as the graph data. Our system directly mimics the prompt engineering of EDUHints to ensure consistency with results. Once the prompt has been built, hint generation can be run locally.

Live Updates

Optionally, we have the ability to classify entities on the fly if they are new to the system, which is done before querying our graph database. We accomplished this by using few-shot prompting with our small language model to classify an entity into our ontology by having it group the entity into its type and subtypes. This process showed promise but when compared to the speed and consistency of using a third party large model such as Openai's gpt-3.5.turbo, it was clear more work would be needed to make this more reliable within our system. We discuss limitations of SLMs in this manner further within section 7.1. When evaluating the performance of the GraphRAG we turn off the option to run live updates.

5. Evaluations/Metrics

We evaluated the two systems, EDUHints and GraphRAG by generating hints for the same set of 25 student logs randomly sampled from classroom data. Each log contained the student's recent bash commands, 3 most recent questions asked, and knowledge-check responses. Knowledge-check responses are the answers to practice questions throughout the exercise. For each log, we produced two hints: one with EDUHints and one with GraphRAG. These paired hints were presented blindly and in random order to a panel of four cybersecurity instructors to minimize bias toward either system.

Each instructor was asked to vote for the hint that they believed was more educationally useful for the student. If both hints were deemed equally good, instructors could select "Similar." In total, this process yielded 100 individual votes (25 logs × 4 voters).

To better understand system performance, we aggregated the votes in two ways:

- **Votes:** the raw number of instructor votes for each system across all 25 logs.
- **Wins:** the number of logs for which a system received the majority of votes (i.e., at least 3 out of 4 votes for that log). If neither system held a majority or if "Similar" was the majority choice, that log was counted as a win for Similar.

6. Results

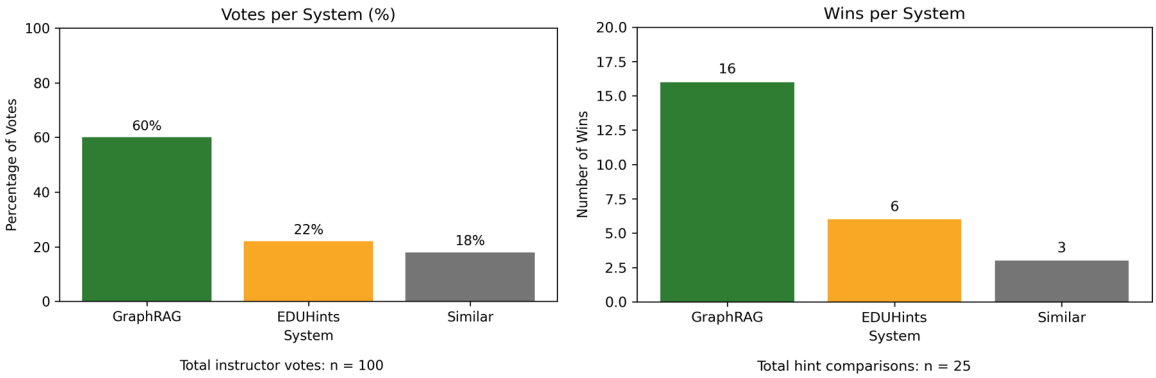


Figure 3: GraphRAG vs EDUHints Results: Figure 3 compares instructor preferences across EDUHints, GraphRAG, and instances where both systems were judged similarly

These results show that our GraphRAG architecture consistently outperformed EDUHints. In both the votes per system and wins per system, a majority of instructors preferred GraphRAG generated hints. While our system was shown to produce more beneficial hints for students, the data there are still areas for improvement.

6.1 Examples

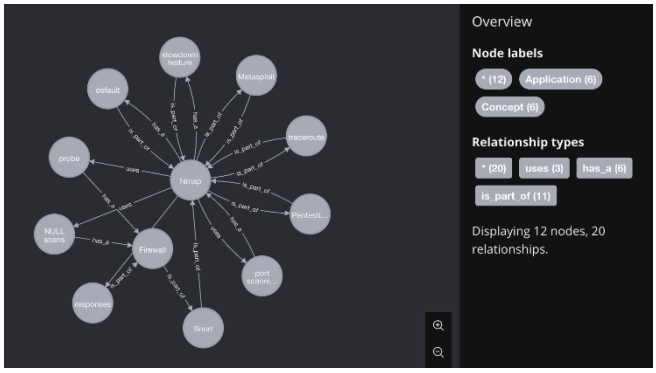


Figure 4: Graphical Output within Neo4j

```
MATCH path = (n)-[*1..2]->(m)
    WHERE
    n.name in ["Firewall", "Nmap"]
    AND
    m.name in ["Firewall", "Nmap"]
RETURN DISTINCT path
LIMIT 10
```

Figure 5: Example Cypher query

Above and to the right, we display an example and a graphical output of a simple Cypher query which finds all distinct paths where either the first or last entity within 2 hops is named "Firewall" or "Nmap".

Below, we show an example of graph data that is injected into our small models context along with a display of which algorithm resulted in that point of data. In this example the student's recent questions asked about viewing a text file and using the ls command.

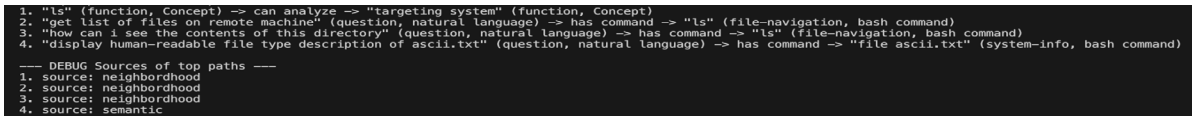


Figure 6: Formatted graph data and sources display

7. Conclusion and Future Work

In this paper we've demonstrated how SLMs with GraphRAG can substantially enhance the quality of hints in cybersecurity education. By leveraging domain specific knowledge graphs, multi-step querying, and semantic scoring, we show that even general purpose lightweight models can learn fundamental concepts of domain specific knowledge. These results underscore the potential for graph-centric retrieval methods that are scalable, privacy-focused, and close the gap between SLMs and LLMs in technical domains.

7.1 Limitations

While instructors clearly favored GraphRAG produced hints, we acknowledge that the small sample size of 25 student logs and 4 cybersecurity instructors limits the overall statistical of our conclusions. Our evaluation strategy as well does not measure the tangible usefulness of hints, rather it aims to measure the perceived usefulness. A possible solution would be to have instructors assess hints, track lesson statistics, and also survey or live prompt students to see which hints they prefer.

Our GraphRAG system is also not entirely local. We utilize a third party LLM to extract cybersecurity relevant entities from student questions or the data we give it. While internal testing was done to use our SLM for entity extraction, this process turned out to be unreliable, resulting in either a reduced amount or irrelevant set of graph data. Right now this was identified as a technical point of limitation but as SLMs continue to improve in the realm of cybersecurity, it is reasonable to suspect the use of third party LLM's will no longer be necessary.

Another limitation we observed was the timeliness in generating hints for the GraphRAG system. Despite practically identical prompt structures for both the EDUHints and GraphRAG, the additional four points of graph data resulted in a significant increase in hint generation time (~10x compared to the EDUHints system). This is likely because the increased context introduced a layer of complexity that the SLM could not handle efficiently. To verify this hypothesis, we as well independently measured the time to query the graph database to identify if this was a bottleneck in our system. Over the 25 hints generated, we observed that querying our database took an average of 3.2 seconds with little deviation, meaning this was not the root cause of the extended time to generate a hint.

7.2 Future Work

In the future we plan to research the following areas:

- **Expanding the instructor-survey study** with a larger panel and more varied student logs to strengthen reliability.
- **Examining cost savings** relative to time saved: Do generated hints result in tangible benefits for teachers, such as time savings .
- **Investigating student-learning metrics** (task-completion time, error-rate reduction) to complement instructor preferences.
- **Refining entity extraction using SLMs**: explore prompt-engineering and guardrails to eventually replace the LLM dependency while maintaining accuracy.
- **Broadening our data**: integrate instructor-authored slides, scenario-specific artifacts, and future EDURange exercises for richer context.

In summary, this work has demonstrated the value of combining SLMs with graph-enhanced retrieval to improve domain-specific hint generation. The progress so far shows that these tools can be made both accessible and effective, and the next phase of research will aim to make them more pedagogically aligned, better validated, and truly end-to-end local, broadening their impact in cybersecurity education and potentially in other technical learning domains

Acknowledgements

This work was partially supported by the National Science Foundation under Awards 2216485 and 2216492.

AI Declaration: AI tools were used to generate the contents of some prototypal scenario context files, to assist in the detailing of figures, and minimal support with code.

Ethics Declaration: The Institutional Review Board of Lewis & Clark College has determined that this project is exempt under exemption code 45 CFR 46.104(d)(1) and will not require further review or approval from the committee.

References

Agrawal, G. et al (2023) "AISecKG: Knowledge Graph Dataset for Cybersecurity Education", Proceedings of the AAAI 2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE), https://par.nsf.gov/servlets/purl/10401616

Chetri, C. (2024) "Exploring Large Language Model-Powered Pedagogical Approaches to Cybersecurity Education" Arxiv, 163-66, https://dl.acm.org/doi/pdf/10.1145/3686852.3686887

Edge, D. et al (2024) "From Local to Global: A Graph RAG Approach to Query-Focused Summarization", ArXiv, https://arxiv.org/abs/2404.16130

Es, S. et al (2024) "Ragas: Automated Evaluation of Retrieval Augmented Generation", Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, https://arxiv.org/pdf/2309.15217

Harjono, K. (2025) "From Embeddings to Entities: A Comparative Analysis of RAG Architectures in Academic Domains", Thesis, https://open.library.ubc.ca/media/stream/pdf/52966/1.0448869/5

Lewis, P. et al (2020) "Retrieval-augmented generation for knowledge-intensive NLP tasks", Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS), pp 9459–9474, https://dl.acm.org/doi/abs/10.5555/3495724.3496517

Liffiton, M. et al (2023) "CodeHelp: Using Large Language Models with Guardrails for Scalable Support in Programming Classes", Proceedings of the 23rd Koli Calling International Conference on Computing Education Research (Koli Calling), pp 1–11, https://doi.org/10.1145/3631802.3631830

Liu, S. et al (2024) "Can Small Language Models With Retrieval-Augmented Generation Replace Large Language Models When Learning Computer Science?", Proceedings Innovation and Technology in Computer Science Education (ITiCSE), pp 388–393, https://doi.org/10.1145/3649217.3653554

Lu, Z. (2025) "Small Language Models: Survey, Measurements, and Insights", https://arxiv.org/abs/2409.15790

Pasquale, R. D. and Represa, S. (2024) "Empowering Domain-Specific Language Models with Graph-Oriented Databases: A Paradigm Shift in Performance and Model Maintenance", https://arxiv.org/pdf/2410.03867

Sahoo, P. et al (2021) "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications", https://arxiv.org/pdf/2402.07927

Sarmah, B. at al (2024) "HybridRAG: Integrating Knowledge Graphs and Vector Retrieval Augmented Generation for Efficient Information Extraction", Proceedings of the 5th ACM International Conference on AI in Finance (ICAIF), https://doi.org/10.1145/3677052.3698671

Westenfelder, F. et al (2025) "LLM-Supported Natural Language to Bash Translation", Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, https://doi.org/10.18653/v1/2025.naacl-long.555

Wolff, T. et al (2025) "EDUHints: A Human-in-the-Loop Small Language Model System for the Generation of Cybersecurity Hints", Proceedings of the 24th European Conference on Cyber Warfare and Security (ECCWS)

Systematic Literature Review: Challenges and Issues in the Adoption of SOAR Technology in Cybersecurity

Turki Lazzam Alshammari and Talal Al-Balawi
Department of Computer Science, College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

444008877@sm.imamu.edu.sa
tsalbalawi@imamu.edu.sa

Abstract: With the increase in the rate of cyber threats, such as ransomware, social engineering, and zero-day exploits, it is urgent to adopt new security mechanisms like Security Orchestration, Automation, and Response (SOAR) systems. The increase in cyber threats has not only amplified in frequency but also in sophistication. This escalation has forced organizations to rethink traditional defense strategies. SOAR has shown itself to be an important solution by automating repetitive tasks and helping security teams in focusing on strategic threat hunting as well as mitigation. The integration of AI and ML in SOAR frameworks helps in predictive analytics, in which systems can anticipate potential breaches based on pattern recognition from vast datasets. The role of blockchain is to enhance data integrity and help enable secure and decentralized threat intelligence sharing between stakeholders. This paper presents a systematic literature review (SLR) on recent advancements in SOAR technologies, especially the incorporation of artificial intelligence (AI), machine learning (ML), and blockchain; it also reviews case studies across various industry sectors, such as healthcare, finance, industrial control systems, and critical infrastructures, as well as the challenges facing SOAR adoption. By examining 29 studies from academic research, industry case studies, and technical reports, the review synthesizes methodologies, architectures, and performance outcomes to summarize the current state of SOAR systems. The research found that SOAR can significantly reduce incident response times and improve threat detection accuracy, with findings indicating that SOAR can lower response times by up to 80% compared to legacy systems, although implementation costs may reach as high as $5 million. Additionally, specialized personnel are still needed to operate these systems. The skills gap increases barriers to adoption, as few professionals possess expertise in cybersecurity as well as in automation tools. Future directions emphasize developing hybrid models that blend human intuition with machine efficiency for more robust defenses. Finally, the review discusses future research directions to help SOAR further scale, interoperate across platforms, and enable autonomous decision-making.

Keywords: SOAR, S-PSS, Security automation, Incident response, Threat intelligence, Cybersecurity

1. Introduction

Continuously changing threats in the cybersecurity landscape are constantly altering defense mechanisms. The trend of interconnected systems in modern enterprises and critical infrastructure sectors is growing, yet a single breach can give rise to cascading consequences. In other words, the development of Security Orchestration, Automation, and Response (SOAR) systems has been a crucial advancement in response to these challenges. SOAR unifies many security tools and processes to automate threat detection, threat response, and the sharing of threat intelligence. The objective is to minimize human error, shorten response times, and thereby improve an organization's resilience to cyberattacks (Waelchli and Walter, 2025). The motivation for SOAR stems from the fact that traditional Security Information and Event Management (SIEM) systems are useful for data aggregation and alert correlation, but the volume of alerts, the number of false positives, and the burden on human analysts remain high (González-Granadillo et al., 2021). With the rising complexity of attacks, cybersecurity teams have to spend more time manually correlating data and ensuring a speedy response. Consequently, SOAR is both desirable and required for modern cybersecurity defense systems (Empl et al., 2022, Mclaughlin and Elliott, 2023).

Recent work on Smart-Product Service Systems (S-PSS) highlights how the cyber-physical integration of products and services creates new security requirements. SOAR platforms can be viewed as a class of S-PSS in which security "products" (tools, playbooks) and "services" (monitoring, incident response) are orchestrated seamlessly. Framing SOAR as an S-PSS helps align this review with service-systems research and underscores the socio-technical aspects of deployment.

The focus of this review is a literature analysis of SOAR, highlighting the technology integration of AI and ML, different industrial sectors affected by SOAR, challenges, and future directions. This SLR seeks to bridge the gap by synthesizing the key findings of 29 studies to provide researchers, practitioners, and policymakers with a comprehensive overview of the present state and future of SOAR technologies. To structure and guide this systematic literature review (SLR), the following research questions (RQs) were formulated:

RQ1: *How are Smart-Product Service Systems (S-PSS), specifically SOAR technologies, being integrated with emerging technologies such as AI, ML, blockchain, and cloud computing?*

RQ2: *What are the key benefits and limitations observed in the adoption of SOAR across various industry sectors, including healthcare, finance, critical infrastructure, and IoT environments?*

RQ3: *What challenges exist in implementing SOAR solutions, particularly regarding cost, integration with legacy systems, and the cybersecurity skills gap?*

RQ4: *What future research directions are proposed to enhance the scalability, interoperability, and autonomy of SOAR systems?*

2. Methodology

As a structured approach to collate and analyze different findings from various sources, this systematic literature review was conducted. The steps of the methodology were as follows:

SLR process	**Identification**	**Search Strings** "Security Orchestration, Automation and Response" OR "SOAR platform*" OR SOAR SOAR AND (AI OR "machine learning" OR "automation" OR "incident response OR blockchain OR "threat intelligence OR IoT)) SOAR AND ("healthcare" OR "finance" OR "SCADA" OR "critical infrastructure")
		Databases Searched: IEEE Xplore, Springer, Scopus, Google Scholar
		Time span: 2019–2025
		Language: English
	Screening Process	Removal of duplicates
		Initial screening based on relevance to SOAR technologies
		Exclusion of pure SIEM/XDR studies without orchestration or automation focus
	Eligibility (Full-Text Assessment)	• Full-text evaluation of remaining studies
		Inclusion of studies addressing: – SOAR platforms or SOAR-adjacent tools – AI/ML, blockchain, and digital twins etc. integration – Particular emphasis on sector-specific implementations (healthcare, finance, ICS, IoT, critical infrastructure) – Empirical evaluation of incident response or threat detection **Exclusion** of non-English, non-peer-reviewed, opinion-based, or non-technical studies
		Total 29 studies included in the review:
	Synthesis and Quality Assessment	Thematic analysis of extracted data
		Bibliometric and keyword co-occurrence analysis (VOSviewer)
		Studies were evaluated based on: (i) clarity of research objectives and system architecture. (ii) methodological depth, (iii) relevance to SOAR adoption or implementation. (iv) presence of empirical evaluation, case study, or performance metrics, and (v) clarity in reporting challenges and limitations.

Figure 1: SLR process, including identification, screening, eligibility assessment, synthesis, and quality assessment

2.1 Literature Search and Selection

Studies related to SOAR techniques were obtained from an extensive search across academic databases, conference proceedings, and technical reports. The identification criteria consisted of the following inclusion criteria: (i) studies that address components of SOAR implementation; (ii) that encompass advanced technologies, such as AI/ML or blockchain; or (iii) studies revealing case studies of sector-specific applications. A total of 29 references that met these criteria were identified (Waelchli and Walter, 2025).

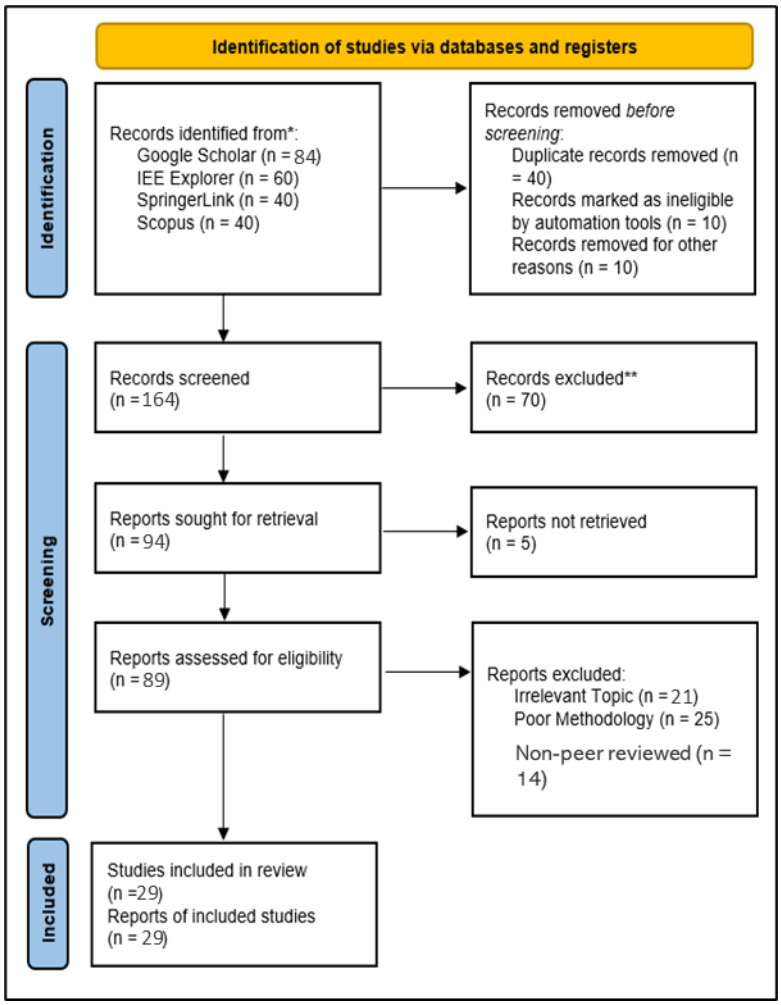


Figure 2: PRISMA flow diagram

2.2 Data Extraction

Information from each study was extracted and condensed from each study (1 study objective, methodology, industry focus, key study results, and challenges identified). Reading the full texts and summarizing the main contributions and limitations of each paper was part of the extraction process (Empl et al., 2022, Christian et al., 2022).

2.3 Thematic Analysis

The green cluster groups AI- and ML-related keywords (e.g., 'machine learning', 'reinforcement learning'), reflecting a strong methodological focus. The orange cluster centers on sector contexts ('healthcare', 'finance', 'ICS'). The red cluster highlights operational themes ('incident response time', 'threat detection'), while the purple and blue clusters denote challenges ('legacy integration', 'skill gap') and future directions ('autonomous SOAR', 'blockchain'), respectively.

The extracted data were grouped into several areas.

- SOAR studies Including ones that integrate AI, ML, and blockchain into SOAR (Alnfiai, 2025, Vinodh Gunnam et al., 2023, Josic et al., 2024)

- Research on SOAR Implementations in health, finance, SCADA systems, and microgrids (Mir and Ramachandran, 2021).
- Incident response times which include measures that assessed the effect of SOAR on incident response time, alert management, and resource optimization (Vast et al., 2021, Gibadullin and Nikonorov, 2021).
- Identified barriers such as legacy integration, cost, and cybersecurity talent shortage (Bhagyalakshmi et al., 2024).

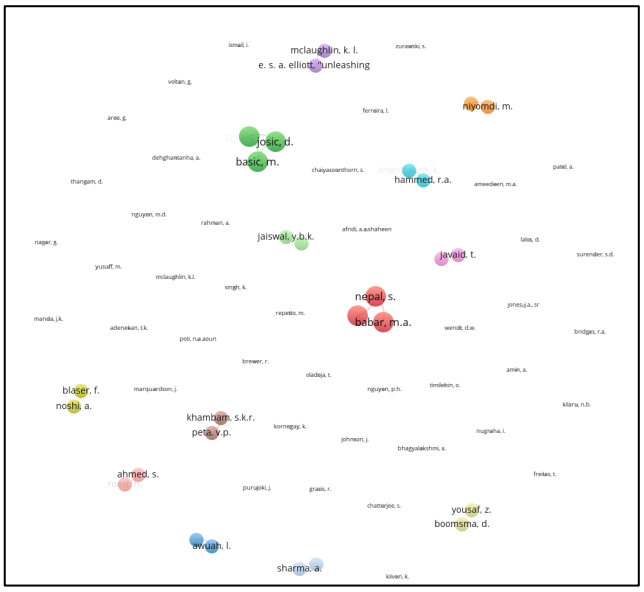


Figure 3: VOSviewer co-occurrence map of author keywords across the 29 studies, with clusters indicating (1) AI/ML integration, (2) Sector-specific applications, (3) Operational efficiency, (4) Challenges, and (5) Future research trends

2.4 Synthesis and Reporting

The final phase was to synthesize the findings from the literature into a coherent narrative. Thematic areas were then used to organize the review, covering a historical perspective as well as a looking ahead to future trends in SOAR.

3. Results

3.1 Evolution of SOAR Technologies

Peer-reviewed literature documents the evolution of SOAR from rule-based automation layers built on top of SIEM systems to intelligent, adaptive platforms that can orchestrate complex, multi-stage incident response workflows. Manual incident response, as well as traditional SIEM platforms, was primarily used in early cybersecurity systems (González-Granadillo et al., 2021). However, as cyber-attacks became more voluminous and complicated, automation became critical to handle overly large and complex data interfaces and to minimize human error. In early research, the focus was on the benefits of data aggregation from multiple sources and the application of basic automation to trigger incident response protocols (Vast et al., 2021, Gibadullin and Nikonorov, 2021). Ultimately, these systems have evolved into sophisticated SOAR platforms that house AI-powered analytics and adaptive response capabilities (Ismail et al., 2025). More recent empirical studies confirm that contemporary SOAR platforms go beyond static playbooks. Bridges et al. (2023) demonstrate that modern SOAR deployments reduce cognitive load on analysts by automating cross-tool data collection and response execution in a significant manner. Their findings highlight that the core value of SOAR is based not just on automation speed, but also encompasses process consistency, auditability, and the reduction of human error in large-scale SOC environments.

3.1.1 Integration with AI and machine learning

As with other trends in SOAR systems, combining AI and ML algorithms is one of the most significant developments in the evolution of SOAR systems. Johnson and Awuah (2021) provide a comprehensive survey of AI-based SOAR architectures. They demonstrate how supervised learning, unsupervised anomaly detection,

and reinforcement learning are being embedded in orchestration pipelines at an increasing rate. These techniques enable SOAR platforms to dynamically prioritize alerts. They also enable the suppression of false positives and adaptation of response strategies based on evolving threat behavior. Vast et al. (2021) have also shown that the use of AI could equip SOAR systems with the ability to reject false-positive findings and focus on detecting true security threats. Nguyen et al. (2024) proposed a multi-layered SOAR architecture that enables RL-based response strategy adaptation in response to real-time threat intelligence. Empirical evaluations also indicate that AI-enhanced SOAR systems function as a force multiplier for security analysts, particularly in environments characterized by high alert volumes. Bridges et al. (2023) report that AI-assisted orchestration substantially improves detection accuracy by correlating contextual indicators across disparate data sources, thereby minimizing redundant investigations. Reinforcement learning approaches further enhance autonomy by enabling systems to learn optimal response actions from historical incident outcomes, reducing reliance on manually curated playbooks.

However, studies differ in the magnitude of results and in their evaluation methods. Vendor or production reports (e.g., Vinodh Gunnam et al., 2023) report large operational gains (20-75% reduction in response time) while laboratory or simulated evaluations (e.g., some RL simulation studies) report promising trends. However, these reports are devoid of real-world validation. Studies indicate a consistent direction (that AI helps), but with variable effect sizes and confidence.

3.1.2 Blockchain and data integrity

Among other things, another emerging trend is the integration of blockchain technology into SOAR platforms. It has been investigated in peer-reviewed literature as a complementary mechanism to enhance trust, integrity, and accountability in SOAR-enabled ecosystems. Recent studies on blockchain-based cyber threat intelligence sharing demonstrate that distributed ledgers can provide tamper-resistant logging, decentralized trust management, and secure information exchange across organizational boundaries. For example, Tolah et al. (2025) proposed a blockchain-enabled threat intelligence framework that ensures the immutability and provenance of shared indicators of compromise. Such properties are particularly relevant to SOAR systems operating in federated environments, where automated response actions depend on externally sourced intelligence. In distributed systems such as microgrids, De Dutta and Prasad (2020) mentioned that blockchain can improve data security and integrity. Creating secure channels between dispersed components of the network using blockchain allows SOAR systems to prevent data tampering and ensures accountability during incident responses. This convergence of blockchain and SOAR is still at an early stage, but there is promise in its use to secure decentralized architectures in cases where there is a high level of trust and transparency (De Dutta and Prasad, 2020, Ismail et al., 2025). However, current studies are largely conceptual or limited to small-scale pilots. No study has demonstrated broad production-scale performance with latency and scalability considerations resolved. Thus, the evidence is promising but of low confidence.

3.2 Sector-Specific Implementations

SOAR has been used globally to solve problems in various sectors, each with its own peculiarities and needs. In this section, we review a number of important studies that target healthcare, finance, industrial control systems, and critical infrastructure (Alnfiai, 2025).

3.2.1 Healthcare

Patient data are highly sensitive, and healthcare systems face many security challenges due to IoT devices, cloud services, and on-premise systems they must manage within their IT infrastructure. Researchers have provided a detailed framework, Security Operations and Automation Platform Architecture (SOAPA), a combination of SOAR and SIEM, which improves cybersecurity in healthcare. The SOAPA framework consists of data services, security operations, and user interface components that provide real-time monitoring and automatic incident response (Chatterjee, 2020).

3.2.2 Finance

Cyber threats are particularly dangerous for the financial sector because sensitive data are of high value, and the applicable laws and regulations are stringent. In financial institutions, Vinodh Gunnam et al. (2023) explored the use of SOAR with AI, which demonstrated systems capable of reducing incident response time to as low as 20-40 minutes in some cases, compared to over 90 minutes in others. The investigation emphasized the advantages of cloud-based SOAR solutions for handling vast amounts of security incidents with high detection accuracy. Although these advantages exist, integrating SOAR with legacy systems and complying

with standards like PCI-DSS and GDPR remain challenges for financial institutions. The integration process often requires significant customization and investment in cybersecurity talent, which can be too costly for smaller organizations (Vinodh Gunnam et al., 2023). These studies show that SOAR offers tremendous value to the financial sector, but its success depends on overcoming integration and regulatory hurdles.

3.2.3 Industrial control systems and critical infrastructure

The most critical data across energy, transportation, and water systems are stored in Supervisory Control and Data Acquisition (SCADA) systems, which are increasingly threatened by sophisticated cyberattacks. SOAR was investigated for its application in SCADA for smart grid systems by Mir and Ramachandran (2021). According to their study, SOAR can reduce high alert volumes and improve response times. However, the integration of IT and Operational Technology (OT) remains a problem since modern cybersecurity protocols were not integrated with legacy SCADA systems (Mir and Ramachandran, 2021). Furthermore, research on microgrids indicates that SOAR can help improve cybersecurity for distributed energy systems. De Dutta and Prasad (2020) demonstrated that SOAR, when combined with AI/ML and blockchain technologies, can form a robust multi-layer defense for microgrids against cyberattacks. However, automating incident responses in microgrids is difficult due to their decentralized nature and the need for continuous updates and adaptations (De Dutta and Prasad, 2020, Ismail et al., 2025, Manda, 2021).

3.2.4 IoT and cyber-physical systems

The attack surface for cybercriminals has expanded with the integration of IoT devices into everyday infrastructure. According to Nguyen et al. (2024), the massive volume of IoT-generated data requires processing, which can be managed by specialized SOAR solutions that handle the output from interconnected devices. Additionally, these frameworks must address a wide range of attack vectors, including DDoS attacks and zero-day exploits. As shown in these studies, AI-driven automation holds strong potential to enhance situational awareness and reduce incident response times in the complex IoT environment (Nguyen et al., 2024, Gilbert et al., 2025). However, integrating disjoint systems and scaling SOAR solutions across large-scale networks remain major challenges (Gilbert et al., 2025, McLaughlin, 2023).

Overall, studies in finance and healthcare generally report measurable operational gains but also emphasize regulatory and legacy constraints. For example, Vinodh Gunnam et al. (2023) and Christian et al. (2022) report clear cost and response-time benefits in financial settings, while Mir and Ramachandran (2021) and Chatterjee (2020) highlight severe legacy/OT integration issues in SCADA. In short, context matters: production environments with modern stacks realize larger gains than legacy-heavy OT contexts.

3.3 Operational Efficiency and Incident Response

SOAR has been shown to be a direct benefit to organizations because several studies have quantified the reduction in incident response times. There are many reports noting that the integration of SOAR platforms can lead to substantial drops in incident response times. Gibadullin and Nikonorov (2021) showed, for example, that an open-source SOAR-based incident management system reduced response time from 34.5 minutes to 12 minutes. In another research study, AI-enabled SOAR in banks resulted in more than a 75% reduction in response time, demonstrating the revolutionary influence of automation on operational efficiency (Vinodh Gunnam et al., 2023). In addition to reducing the risk of protracted security breaches, the reduction of the analysis burden on human analysts frees them for more routine work (Gibadullin and Nikonorov, 2021, Vinodh Gunnam et al., 2023). Bridges et al. (2023) also report substantial reductions in mean time to detect (MTTD) and mean time to respond (MTTR), attributing these gains to automated evidence aggregation and response execution across integrated security tools.

Advancements in the accuracy of detecting threats have been made by incorporating AI and ML in SOAR systems. Bhagyalakshmi et al. (2024) and Jaiswal et al. (2024) employed machine learning algorithms to scrutinize large datasets and recognize true and false positives with high accuracy. For example, AI-based SOAR systems in phishing attack scenarios have achieved up to 98% protection accuracy to improve the general protection posture of an organization (Vinodh Gunnam et al., 2023; Jaiswal et al., 2024). However, these advancements not only help with the management of alerts but also minimize the burden on security teams that have to manage huge numbers of false alarms - something that is particularly important in large-scale network environments (Vast et al., 2021). This is particularly important for organizations that have an enormous number of security logs and alert feeds per day (Vast et al., 2021, Gibadullin and Nikonorov, 2021).

Another crucial benefit from SOAR implementation is operational cost savings. Christian et al. (2022) demonstrate that a cloud-native SOAR solution can scale to handle over 10,000 security incidents a month for under $65 per month. Such a low-cost solution provides a viable pathway for small and medium-sized enterprises (SMEs) that could not afford the high-end commercial solutions. Nevertheless, as these studies demonstrate promising results with regard to cost efficiency, there still remain significant initial investments in training and system customization (Christian et al., 2022).

Moreover, there is also inconsistency in the reporting of results. Open-source deployments reported by Gibadullin and Nikonorov (2021) showed response-time reduction (34.5 → 12 minutes), whereas other reports show percentage improvements without baseline clarity. Differences arise from inconsistent outcome reporting (absolute minutes vs % reduction vs qualitative gains).

3.4 Integration Challenges and Limitations

Although the benefits offered by SOAR technologies are clear, many studies show a reluctance to advance these technologies. Broadly speaking, these challenges can be broken into technical, operational, and human factors. Among the most mentioned challenges across several studies is integrating SOAR with legacy systems. One issue is that many organizations have a mix of modern and older systems, and this makes it harder to combine these components with SOAR, making integration into a heterogeneous environment quite a technical obstacle. For instance, Roche and Dowling (2023) mention that the integration of SOAR with existing SIEM systems, e.g., Splunk, generally requires a lot of customization to incorporate legacy workflows and automatic rules. Similarly, sectors like healthcare and finance heavily rely on data interoperability, and this integration complexity becomes even greater with the presence of legacy systems (Vinodh Gunnam et al., 2023). The studies reviewed also suggest that high investments in the upfront collarization of the system, training, and infrastructure upgrades are commonly needed. with even more difficulty for SMEs (Roche and Dowling, 2023). Also, the need for continuous upgrades and maintenance to respond to changing threats puts further financial pressure (Żurawski et al., 2025).

One of the most consistently reported limitations in peer-reviewed SOAR research is the cybersecurity skills gap. While SOAR automates execution, it does not eliminate the need for expert analysts capable of designing playbooks, validating AI outputs, and supervising automated responses. The problem has been cited by many studies including those by Waelchli and Walter (2025) and McLaughlin (2023). Further, SOAR systems face scalability challenges in highly dynamic environments where many distributed nodes exist, such as IoT networks and critical infrastructures. These socio-technical challenges give weight to the consensus that SOAR should be implemented as a decision-support and orchestration layer, rather than as a fully autonomous replacement for human analysts.

3.5 Future Directions and Emerging Trends

Based on the literature review, several research directions are proposed. The first one is the future development of more autonomous SOAR systems that require less human involvement. Nguyen et al. (2024) demonstrate that future SOAR architectures could dynamically adapt their response strategies in response to evolving threat landscapes (Nguyen et al., 2024). Another important research direction specified is the improvement of interoperability between SOAR systems and existing legacy infrastructure using novel middleware solutions. This is an important area of research for sectors where legacy systems are deeply entrenched and replacing all the systems at once is not an option. There is fertile ground for innovation when SOAR converges with other technologies like blockchain, IoT, and digital twins. For example, it can be applied not only to ensure data integrity but also to promote peer-to-peer threat intelligence sharing among various organizations (De Dutta and Prasad, 2020, Ismail et al., 2025). Moreover, digital twin technologies could support stress testing of SOAR playbooks in simulated environments, as well as train AI models in a controlled setting (Ismail et al., 2025). Such integration might additionally enhance the adaptability and robustness of SOAR systems. Thus, the peer-reviewed literature identifies a clear trajectory toward adaptive, learning-enabled SOAR systems that continuously refine response strategies based on operational feedback. Reinforcement learning-based orchestration, standardized integration interfaces, and simulation-based testing environments are repeatedly cited as promising research directions. Johnson and Awuah (2021) emphasize the importance of explainable AI to improve analyst trust and transparency in automated decision processes.

4. Discussion

While the reviewed literature largely agrees that SOAR improves operational outcomes, there are apparent contradictions with respect to effect size and feasibility, which can be reconciled by considering study context

and methods. Vendor or production reports often present larger gains because they report metrics from mature deployments, whereas academic or pilot studies typically evaluate prototypes or simulated workloads and therefore report smaller or preliminary gains. Differences in metrics (absolute MTTR in minutes, percentage reduction, or qualitative analyst workload reduction) further complicate direct comparisons. Consequently, practitioners should interpret large, single-study effect estimates cautiously and prefer conclusions triangulated across multiple high-quality studies.

Moreover, reviewed literature suggests that complete automation has not been achieved. Human oversight is still present and essential for the situations that involve ambiguous, new, or context-heavy threats that automated playbooks cannot handle still. Researchers have shown that automated systems do not have the interpretive capability and the knowledge of domain. They also lack contextual judgment which is required for complex attacks. It means that analysts are still responsible for preventing automation-based errors and for making sure that there is safe decision-making. The literature also warns that over-reliance on automation introduces issues such as misclassification, reduced transparency, and challenges in post-incident analysis. This issue further supports the need for hybrid human-machine collaboration (Mclaughlin and Elliott, 2023, Żurawski et al., 2025, Nugraha, 2021). Furthermore, researchers recommend that there must be scalable adoption models and phased deployments, which are considered more sustainable pathways. Further, cost-benefit analyses show that reductions in response times and mitigation of breach-related losses can justify investment when weighed against tangible as well as intangible organizational benefits (Vinodh Gunnam et al., 2023, Christian et al., 2022). Overall, the reviewed literature indicates that the evolution of SOAR depends not only on technical innovation but also on improved human - machine alignment. It particularly depends on sustainable cost models and more seamless integration strategies that are able to support different kinds of organizational ecosystems (Kilaru et al., 2021, Islam et al., 2020, Nugraha, 2021).

5. Conclusion

We synthesized findings from 29 studies related to SOAR technologies to create this systematic literature review of the state of the art of SOAR. The review discusses how SOAR systems have progressed from simple automation tools to complex platforms that leverage several tools such as AI, ML, and even blockchain for enhancing cybersecurity operations in various sectors. The balance of evidence supports that SOAR improves incident handling and detection accuracy, but heterogeneity in evaluation methods, sectoral contexts, and reporting metrics prevents precise, generalizable effect-size estimates. Where contrasting results appear, they can largely be explained by differences in context, measurement, and maturity of deployment. This review acts as a wake-up call for both academic researchers and industry practitioners to capitalize on the full power of SOAR to secure our shrinking digital world.

AI declaration: No AI tools were used in the preparation of this manuscript. All content was developed solely by the authors.

Ethics declaration: The study is entirely based on previously published literature; therefore, no institutional ethical approval was required.

References

Alnfiai, M.M. (2025) “AI-Powered Cyber Resilience: A Reinforcement Learning Approach for Automated Threat Hunting in 5G Networks”, EURASIP Journal on Wireless Communications and Networking, 2025, 68.

Bhagyalakshmi, A., Laya, C.S., Preethikaa, A.Y. and Varsha, V. (2024) “Machine Learning Based Early Detection of Ongoing Cyber-Attacks”, 2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), IEEE, pp 766-771.

Bridges, R. A., Rice, A. E., Oesch, S., Nichols, J. A., Watson, C., Spakes, K., ... & Erwin, S. (2023). Testing SOAR tools in use. Computers & Security, Vol 129, 103201.

Chatterjee, S. (2020) “Using SIEM and SOAR for Real-Time Cybersecurity Operations in Oil and Gas”, International Journal of Innovative Research and Creative Technology, Vol 6, pp 1-11.

Christian, J., Paulino, L. and de Sá, A.O. (2022) “A Low-Cost and Cloud Native Solution for Security Orchestration, Automation, and Response”, International Conference on Information Security Practice and Experience, Springer, pp 115-139.

De Dutta, S. and Prasad, R. (2020) “Cybersecurity for Microgrid”, 2020 23rd International Symposium on Wireless Personal Multimedia Communications (WPMC), IEEE, pp 1-5.

Empl, P., Schlette, D., Zupfer, D. and Pernul, G. (2022) “SOAR4IoT: Securing IoT Assets with Digital Twins”, Proceedings of the 17th International Conference on Availability, Reliability and Security, pp 1-10.

Gibadullin, R. and Nikonorov, V. (2021) “Development of the System for Automated Incident Management Based on Open-Source Software”, 2021 International Russian Automation Conference (RusAutoCon), IEEE, pp 521-525.

Gilbert, C., Gilbert, M.A., Dorgbefu, M., Leakpor, D.J., Gaylah, K.D. and Adetunde, I.A. (2025) “Enhancing Detection and Response Using Artificial Intelligence in Cybersecurity”, International Journal of Multidisciplinary Research and Publications (IJMRAP), Vol 7, pp 87-104.

González-Granadillo, G., González-Zarzosa, S. and Diaz, R. (2021) “Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures”, Sensors, Vol 21, 4759.

Gunnam, V., Cheemakurthi, S.K.M. and Kilaru, N.B. (2023) “And PCI Data Security with Cloud Innovations”, International Journal of Advances in Engineering and Management, Vol 5, pp 974-980.

Islam, C., Babar, M.A. and Nepal, S. (2020) “Architecture-Centric Support for Integrating Security Tools in a Security Orchestration Platform”, European Conference on Software Architecture, Springer, pp 165-181.

Ismail, Kurnia, R., Brata, Z.A., Nelistiani, G.A., Heo, S., Kim, H. and Kim, H. (2025) “Toward Robust Security Orchestration and Automated Response in Security Operations Centers with a Hyper-Automation Approach Using Agentic Artificial Intelligence”, Information, Vol 16, 365.

Jaiswal, R., Marshal, R., Rao, V.V. and Singh, K.P. (2024) “AI Phishing Detection Framework for Businesses with Limited Resources”, 2024 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT), IEEE, pp 399-404.

Josic, D., Basic, M. and Zgrablic, L. (2024) “Security Principles in Cloud Computing”, Annals of DAAAM & Proceedings, 35.

Kilaru, N.B., Cheemakurthi, S.K.M. and Gunnam, V. (2021) “SOAR Solutions in PCI Compliance: Orchestrating Incident Response for Regulatory Security”, ESP Journal of Engineering & Technology Advancements, Vol 1, pp 78-84.

Kinyua, J., & Awuah, L. (2021). AI/ML in Security Orchestration, Automation and Response: Future Research Directions. Intelligent Automation & Soft Computing, Vol 28, 2.

Manda, J. (2021) “Cybersecurity Automation in Telecom: Implementing Automation Tools and Technologies to Enhance Cybersecurity Incident Response and Threat Detection in Telecom Operations”, Advances in Computer Sciences, 4.

McLaughlin, K.L. (2023) “Defense Is the Best Offense: The Evolving Role of Cybersecurity Blue Teams and the Impact of SOAR Technologies”, Edpacs, Vol 67, pp 35-41.

McLaughlin, K.L. and Elliott, E.S. (2023) “Unleashing the Power of Mobile Threat Hunting Toolkits: Why They Are Crucial in Today’s Cybersecurity Landscape”, EDPACS, Vol 68, pp 1-6.

Mir, A.W. and Ramachandran, R.K. (2021) “Implementation of Security Orchestration, Automation and Response (SOAR) in Smart Grid-Based SCADA Systems”, Sixth International Conference on Intelligent Computing and Applications: Proceedings of ICICA 2020, Springer, pp 157-169.

Nguyen, M.-D., Mallouli, W., Cavalli, A.R. and Montes de Oca, E. (2024) “AI4SOAR: A Security Intelligence Tool for Automated Incident Response”, Proceedings of the 19th International Conference on Availability, Reliability and Security, pp 1-8.

W., Cavalli, A.R. and Montes de Oca, E. (2024) “AI4SOAR: A Security Intelligence Tool for Automated Incident Response”, Proceedings of the 19th International Conference on Availability, Reliability and Security, pp 1-8.

Nugraha, I. (2021) “A Review on the Role of Modern SOC in Cybersecurity Operations”, Int. J. Current Sci. Res. Rev, Vol 4, pp 408-414.

Roche, D. and Dowling, S. (2023) “Elevating Cybersecurity Posture by Implementing SOAR”, 2023 Cyber Research Conference-Ireland (Cyber-RCI), IEEE, pp 1-7.

Tolah, A. (2025) "BlockIntelChain: a blockchain-based cyber threat intelligence sharing architecture", Scientific Reports. https://doi.org/10.1038/s41598-025-29152-6

Vast, R., Sawant, S., Thorbol, A. and Badgujar, V. (2021) “Artificial Intelligence Based Security Orchestration, Automation and Response System”, 2021 6th International Conference for Convergence in Technology (I2CT), IEEE, pp 1-5.

Waelchli, S. and Walter, Y. (2025) “Reducing the Risk of Social Engineering Attacks Using SOAR Measures in a Real World Environment: A Case Study”, Computers & Security, Vol 148, 104137.

Żurawski, S., Chrzyszcz, A., Ciekanowski, Z., Pauliuchuk, Y., Pietrzyk, S. and Wyrzykowska, B. (2025) “Effectiveness of Information Security Incident Management Systems: Identifying Practices, Challenges and Development Perspectives”. European Research Studies Journal, 2025.

The Cyber Physical Immune System: A Self-Healing V2X Framework for Global Defense and Healthcare

Babajide Asaju[1,2], Almustapha Wakili[1], Saugat Guni[1] and Woosub Jung[2]
[1]Towson University, USA
[2]Naval Postgraduate School, USA

basaju@towson.edu (corresponding author)
awakili2@towson.edu
sguni1@towson.edu
woosubjung@towson.edu

Abstract: The cyber-physical immune system (CPIS) is an important field in healthcare security, a chemistry reactor-based, AI-friendly framework that shields healthcare and urban structures like the human immune system. This self-sufficient network identifies a threat, contains it, and neutralizes it in minutes without any disruption to company operations. John Arquilla (2021) coined this era of social networks and" netwars." Cyberwarfare is on the rise around the world, and individuals, commercial companies, and institutions of all kinds are among the first to feel its sting. AI and real-time CPIS analytics allow for the expansion of coverage based on new threats, thus connecting cybersecurity and critical infrastructure protection. These ideas make it possible to identify the conceptual background, main components, and transformative characteristics of CPIS, as well as to consider its potential use in V2X communications, worldwide defense and medicine to overcome difficulties and facilitate further development. In this paper, we present the current threats in research with the prevailing challenges of current self-healing in CPIS for both global defense and healthcare systems. We explain the key components of the self-healing functionality of the CPIS in healthcare.

Keywords: AI-driven systems, Smart medical devices, Cybersecurity

1. Introduction

The current climate calls for an urgent need for a holistic investigation of the challenges that overwhelm the defense system and healthcare infrastructure. Every minute, numerous threats loom over the Internet of Things (IoT) and Cyber-Physical Systems (CPS) devices across all infrastructures. Within the IoT and CPS domains, many studies have been conducted on vehicle-to-everything (V2X) communications to help different forms of transportation applications for road safety, traffic accuracy, user threat information, and patient privacy. This technology facilitates direct communication between vehicles (V2V) and between vehicles and the traffic infrastructure (V2I) Rammohan et al. (2023). The proliferation of V2X communication systems raises fundamental questions regarding the protection of individual privacy rights Asaju (2024). Meanwhile, the digital environment has introduced many new threats, especially in the public and private healthcare infrastructure and systems that are vital to V2X communications and health security.

With the new and growing complex world of technology, there is a growing demand for sound cyberphysical systems. The combination of physical infrastructure, communication systems, and smart algorithms has introduced a new class of systems called cyber-physical systems (CPS), which have transformed various fields, including healthcare, transportation, and smart cities. However, the integration of these systems has exposed them to various cyber-risks, including malicious, accidental, and unintended disruptions. Due to this challenge, a novel technology called a cyberphysical immune system (CPIS) has been introduced as a fashionable solution Ferrag et al. (2018).

This study adopts a narrative research method within the qualitative methodology, using existing literature to highlight theories and machine learning algorithms to adapt AI and blockchain to cybersecurity in the healthcare and transportation sectors. The medical infrastructure must be robust against external threats and attacks due to the sensitive and personal information they collect Gatouillat et al. (2018). As these technologies advance, there is a need to expand security mechanisms to protect connected systems, such as medical facilities, from emerging risks. The study focuses on improving the scalability of these technologies and reducing their high computational requirements, which will open possibilities for their use. This study explores the complexity of cyber-physical information systems (CPIS) in the Internet of Things (IoT) and vehicle-to-everything (V2X) communication, examining potential opportunities for efficiency and functionality, especially in sensitive domains such as health and transportation. V2X technology is expected to offer various services to create a safer and more comfortable user environment Liu et al.(2020b).

To address these challenges, researchers have explored CPIS as a viable solution derived from the natural human immune system. This approach can detect, isolate, neutralize, and recover from cyber threats. The literature reveals that techniques have been developed to identify and control threats as they arise. In addition, blockchain technology has been investigated for its potential to provide enhanced data security and faultless messaging in distributed systems, helping to prevent advanced threats Morris et al. (2009). The study also examines various privacy-preserving schemes that have been suggested in the literature. Attaining a solution that satisfies the requirements of low communication overhead, real-time processing, and a high level of security while maintaining a lightweight size that can be deployed on edge devices is essential Boubaker et al. (2023). A side-by-side comparison of different security and privacy approaches for fog-based IoT medical applications was performed to evaluate their effectiveness in protecting sensitive information in healthcare settings Ponnambalam (2010).

Although CPIS and blockchain technologies offer promising solutions for improving cybersecurity in healthcare and transportation, their high computational requirements and scalability issues currently limit their widespread implementation in sensitive domains, such as medical facilities and V2X communication systems. The Morris water maze test is a widely used behavioral task to assess spatial learning and memory in rodents. This test involves placing an animal in a circular pool filled with opaque water, where it must locate a hidden platform using spatial cues.

Figure 1 presents an example of the visualization of medical-based nodes within the framework of a comprehensive V2X CPIS. The model is an integral component of an interconnected healthcare security architecture designed to detect cyber-physical threats in real time, thereby enabling responsive threat management.

The purpose of this study encompasses the following key areas that aim to contribute to improvements in this field.

- The main goal of this study is to recognize the latest machine learning tools, methods, and algorithms to interceptframework for global defense and healthcare self-healing functionality into cyber–physical immune systems.
- The capability of cyber-physical immune systems will be analyzed with respect to state-of-the-art technology, and artificial intelligent tools and methods in carrying out the overall functionality of self-immunology will be explored.
- The current literature will be critically reviewed to identify current state of techniques, tools, algorithms, modelclassifications, methods, frameworks, networks, and architectures currently deployed for a cyber-physical immune approach.

The remainder of this paper is organized as follows. Section 2 provides a theoretical framework for the development of AI-based cyber-sensitivity surveys on how threats can be detected and confined. Section 3 presents an overview of IoT and V2X using Machine Learning Background, structures, and typical applications in the global defense healthcare infrastructure. Section 4 presents an overview of the background, machine learning, and blockchain applications of cyber-physical immune systems (CPIS) and their applications in healthcare infrastructure. In Section 5, we discuss how the protection of IoT and V2X using machine learning and blockchain in cyberphysical systems can have a greater impact on real-time threats to immune systems. Some significant open issues and research challenges for the Integration of ML and Blockchain in Cyber Security and much larger perspectives are addressed. In Section 6, we address how the scalability and cost of the solutions can be used to protect patient data privacy in the blockchain domain. Finally, we conclude this paper on the advancement of cyber-physical systems development of newer methods to make AI work easier to interpret data protection in Section 8.

2. Theoretical Framework for the Development of AI-Based Cyber Security

The cyber-physical immune system (CPIS) is inspired by the concept of a human immunological system, which entails the ability to identify threats, confine them, and eliminate them to preserve systemic health. In biological systems, the immune system operates through two primary layers: the natural defense, which is rapidly general, and the acquired defense, which becomes more specific over time Polo (2024). Therefore, a double-sorted strategy is necessary to prevent the development of new methods to avoid protection. Likewise, CPIS incorporates these ideas into a theory of cyber security and uses biological approaches to protect the infrastructure against different types of cyber threats.

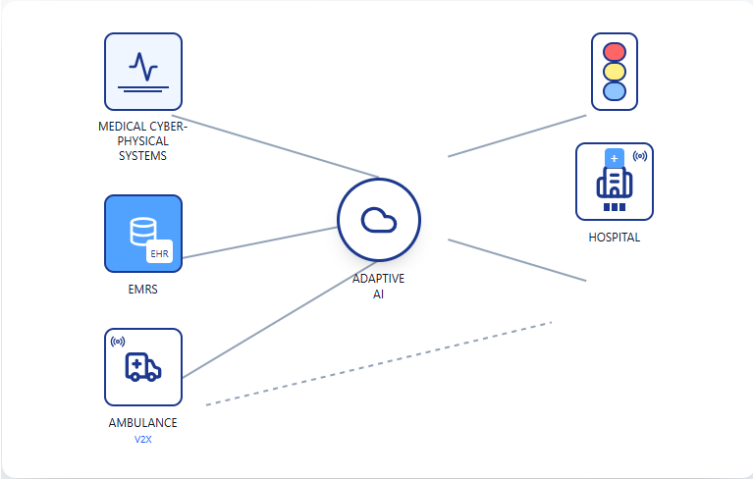


Figure 1: Simulation Visualization of a Hospital Based Node

In the human immune system, pathogen recognition is performed by pattern recognition receptors that grab foreign antigens that stick to body cells and create defensive mechanisms. This concept is also seen in CPIS, and in the system, the AI algorithms serve the purpose of the detection component, which monitors network activity and identifies signs of malicious activity.

In CPIS, isolation entails cordoning off affected nodes in a network to avoid the extension of threats to other parts of the network. More specifically, Farooq and Zhu (2020) noted that network segmentation could be a valuable way to prevent cyber risks, as well as using concepts reminiscent of biological containment. Neutralization is the last reaction to threats that exist in the environment, through specific actions carried out on objects. Similarly, in CPIS, AI countermeasures neutralize threats, after which the system quickly restores operations Hasan et al. (2020), and provides information on secure and reliable recovery concepts in cyber-physical domains to achieve operational continuity, such as the body's ability to work after defeating an infection.

AI is central to the CPIS and is responsible for the defense adaptability of the system. Such algorithms allow CPIS to analyze vast quantities of raw data, which, when analyzed, will show signs of possible threats that are in their early stages. Machine learning models could be extremely good at discovering new and previously unknown use cases of IoT, which is quite important for the complexity of CPIS. The ability of AI to learn means that CPIS continues to be potent when marked threats change, in the same way adaptive immunity fortifies itself with lessons from previous exposure. Chen *et al*. (2020) showed that another positive change element is self-healing which is helped by AI to allow CPIS to react and patch up an attack with barely any input from people. With embedded real-time analysis, the system can quickly assess damage and team responses and formulate a response. Chaterji *et al.* (2018) highlighted the importance of such capabilities to protect themselves from more advanced cyberattacks. At the same time, Bajpai *et al.* (2020) noted that real-time decision management enables us to minimize risks and maximize response effectiveness and efficiency.

As shown in Figure 3, the CPIS cybersecurity framework was designed as a voluntary risk-based framework to improve cybersecurity in critical global defense healthcare infrastructure. Credit goes to the former President Barrack Obama's executive order calling for the establishment of a brand-new set of standards, guidelines, and practices to help organizations charged with providing the nation's financial, energy, healthcare, and other critical systems to better protect their information and physical assets from cyberattacks Ajakwe et al. (2022).

As observed in deliberations regarding IoT security threats, machine learning is increasingly being used in analytics and countermeasure detection Weingart (2000).

To ensure safety and optimize efficiency, it is imperative to establish a connection between vehicles and external systems through secure data exchanges. If these systems are not reinforced with secure mechanisms, they can be attacked and the results can be horrendous. Studies in this field have described different measures, including cryptographic solutions, safe network connections, and additional applications of intelligent technologies for the real-time discovery of inconsistencies [Rao et al. (2023)].

3. Protecting IoT and V2X using Machine Learning and Blockchain

Modern computing machine interactions can take on different forms of remote connectivity to IoT environments, where remote computing can function as a standalone microsystem by ensuring the security of V2X communications, which is essential to facilitate access to network resources. Various solutions have been proposed to address the security challenges most common to blockchain applications, which can be classified into cryptography-based and trust-based schemes. Compared to the data, Hakeem et al. Hakeem et al. (2020) suggested that authentication and preservation of privacy are among the most pressing concerns in V2X communication. Their application can be seen in different industries, protecting them from future cyber threats.

3.1 Healthcare Systems

The concept of CPIS plays an important role in protecting defense technologies from cyber threats. Emerging healthcare systems interact with critical medical networks, autonomous systems, and important infrastructures, which involve defense threats and require protection against cyber threats that can potentially cause interference or threaten national security. The CPIS can then observe these channels and block the threats identified in system exposure [Dorn (2023)]. Self-healing is more critical in cyber warfare environments, where attackers can proactively and strategically launch precise attacks on important infrastructures such as drones, unmanned aerial vehicles (UAVs), and defense communication networks.

3.2 Healthcare Infrastructure

Several IoT-based systems have permeated the healthcare industry, including remote health monitoring devices, wearable health monitoring devices, and IoT-connected imaging systems. They gather and send patient information, which, to some extent, is pertinent and, therefore, a target for cyber attackers. A threat to cyber security could result in poor patient care, stolen or lost medical records, or the failure of essential devices. CPIS acts as a workaround for these effects by constantly monitoring these devices and protecting patient data from intruders, Das et al. (2021)]. The self-healing capability of the system means that it can identify threats on unique devices, quarantine, and eliminate the threat without interrupting healthcare services. For example, a hospital can protect its medical device system against ransomware attacks that can inevitably paralyze its functions. Thus, CPIS can also help to keep the infrastructure of healthcare facilities secure and functioning during severe attacks, including DDoS, as it can detect the source of the attack and redirect traffic.

3.3 Urban Infrastructure and V2X Frameworks

Due to the actively developing concepts of smart cities, highly intelligent cars, and interconnected urban components, questions arise about cyber security. Urban areas have become dependent on digital systems for traffic control, electrical supply, waste removal, and security. The operational complexity has increased because today's autonomous vehicles (AV) and vehicle-to-everything (V2X) communication systems require real-time data exchange between vehicles, infrastructures, and other smart sensors. The former can improve the reliability of urban critical infrastructure by safeguarding V2X communication networks, which is paramount for the safety and efficacy of driverless vehicles. These networks may also be an open door for hackers and information technology criminals, which can cause crashes or traffic interference [Kiela et al. (2020)].

4. Protecting IoT and V2X using Machine Learning and Blockchain in Cyber-Physical Immune Systems

CPIS mechanisms were developed to protect complex systems from cyber threats, like human immune systems. These systems are extremely important for the Internet of Things and communication between cars and everything else. CPIS uses machine learning and blockchain to improve information security against cyber threats. CPIS combats cyber threats by learning and changing, like the defenses of the body. This is a major step in cybersecurity due to the blending of biology and advanced technology, which is extremely helpful for connected systems and essential services. (W Yu et al., 2025). The IoT also plays an important role in business. IoT has been reported to be one of the most important technologies that will affect US interests in 2025. It defines why organizations seek to improve their ISO 27001 ISMS. These reasons include looking for a way to provide proof of activities, due care, due diligence, and regulatory compliance.

4.1 Real-time Threat Detection and Mitigation

Mitigation in risk management and disaster preparedness encompasses two key objectives: reducing exposure and minimizing possible loss. Each of these categories can contribute to the enhancement of early warning

capability systems. Likewise, infrastructure protection can be strengthened to withstand the potential impacts that can occur with implementable response plans.

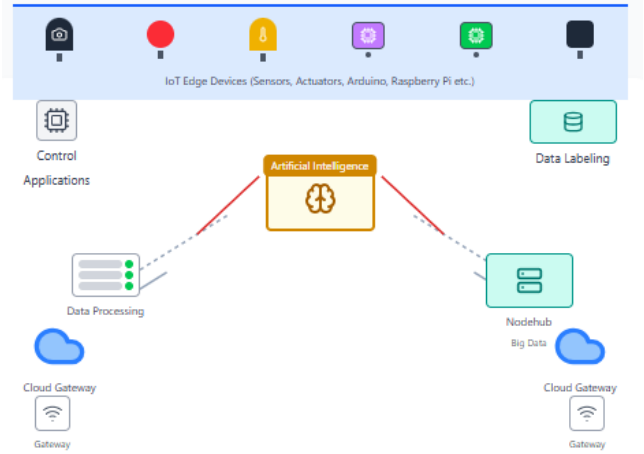


Figure 2: Cyber-Physical Security for IoT Networks

In addition, machine learning can continuously monitor systems to detect any behavior that does not match the typical trends or symptoms of a cyber-attack. This capability enables detection to be performed in real time, providing authorities with the opportunity to intervene before an attack is performed. Automation of Security Responses

The second major benefit of using AI-powered machine learning algorithms in CPS and IoT security is automation. It can provide immediate defensive actions that can be executed when a threat is recognized, such as quarantining a certain component or blocking certain traffic patterns without human intervention. This response minimizes impact and maintains system integrity, which is vital in real-time systems such as health or transportation, where system downtime has disastrous implications [Pandey and Misra (2016)].

4.2 Reduction in System Downtime and Operational Disruptions

Using automated, real-time responses are worked on, thereby lowering the chances of extended system offline time. These systems can bounce back from an attack or several disruptions by reproducing data and rerouting circulation efficiently, thus incurring minimal damage. This benefit is especially valuable in industries where end-user uptime is critical, particularly in healthcare, where interruptions can directly cause patient harm [Farooq and Zhu (2020].

4.3 Improved Resilience to Evolving Threats

Blockchain enhances the reliability of distributed systems in connected environments, proving the reliability of data and network connections. For example, its structure is decentralized, which eliminates the problem of broken control. Due to decentralization and the impenetrable characteristics of the blockchain, no one can manipulate the data, which makes it quite secure and easily checked after a while Rao et al. (2023). This makes it an ideal solution for encrypting sensitive data stored in various systems that must meet the highest trust requirements, including IoT systems in the healthcare and transport industries. In the context of vehicle-to-everything (V2X), the blockchain guarantees that only genuine information is allowed for vehicle and infrastructure data exchange; it prevents the interference of false data from malicious actors, traffic signals, vehicle status, or navigation data [Catrina et al. (2020)].

5. Challenges for the Integration of ML and Blockchain in Cyber Security

The application of machine learning and blockchain in existing protective cyber security systems is not only challenging but also faces several technical and organizational barriers. ML can optimize blockchain networks by improving consensus mechanisms, reducing computational overhead, and facilitating intelligent smart contracts for efficient transaction validation Liu et al.(2020a). Current applications in areas such as health, transport, and finance have not been developed with the input/output requirements of today's technologies, such as blockchain and machine learning. In turn, these systems can be incompatible with advanced solutions, encountering compatibility problems, slowness, and additional costs along the way Zhang et al. (2019).

Blockchain is very advantageous in terms of improving cybersecurity due to its distributed and decentralized properties, such as offering an efficient means to make secure and transparent transactions. However, the use of technology entails a high storage and processing capacity for the database.

5.1 Biased Risks in the Decisions Taken by AI

In the case of AI systems in cybersecurity and healthcare, which are based on data used for training, this problem is rather acute. Poor data used while developing AI models lead to the development of biased models, negatively leading to either insecurity or poor health service delivery. In cybersecurity, for example, if the training data include a single type of threat, the resulting model will also focus on this type of threat and will not notice new or regional ones. This could open systems to cyber-attacks, taking advantage of previously untouched threats. For example, an AI model trained using attack data from one region or a specific network environment is unlikely to identify emergent threats in a different region or a different technological environment of operation from the same threat-actor group.

This also applies to the type of data collected, where certain forms of attack are overly representative, while others receive little representation.

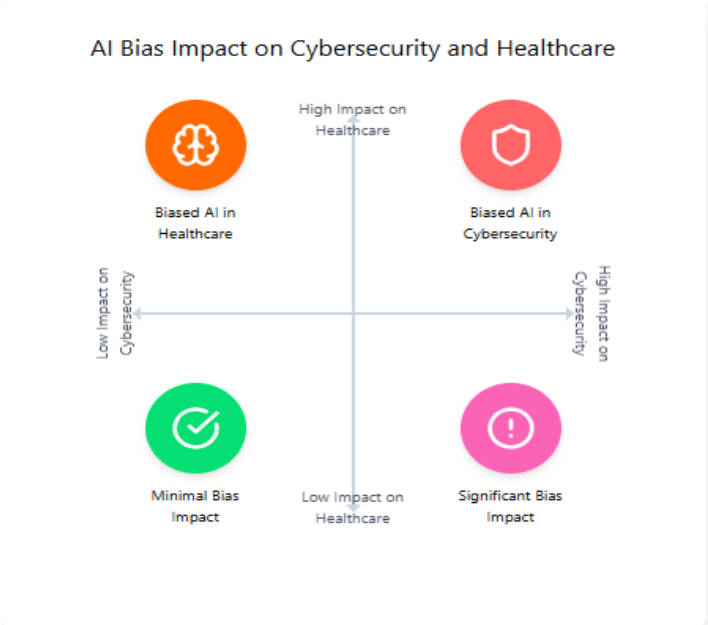


Figure 3: AI Bias Impact on Cybersecurity and Healthcare

A diagram showing the impact of AI bias on cybersecurity and healthcare. which further exacerbates health inequalities. In addition, powered estimation models might be counterintuitive to professional healthcare providers' estimates and might pose ethical issues. The AI system may reproduce bias if it learns from the data that a disparity existed in previous healthcare or security structures. In such cases, AI models use preferential bias, for example, in favor of a certain group, thus reducing fairness in the distribution of items such as healthcare and security.

5.2 Solving Scaling and Cost Problems

The popularity of machine learning and blockchain raises the problem of cost and scalability for many organizations that want to implement these technologies into their security systems. Owing to its decentralized nature, each participant in the blockchain network must process, store, and verify transactions, increasing the rate of storage and computation as the complexity of the network increases Alghamdi et al. (2024). Because every node in the network is required to replicate the entire blockchain ledger and is involved in the consensus mechanism, the cost of the infrastructure increases, preventing the effective massive deployment of blockchain solutions. This is even more difficult in sectors with huge networks of connected devices, such as the health and transport sectors, where every device and system must be secure.

5.3 Scalability and Costs of Solutions

The use of AI and blockchain in sensitive domains such as medicine and warfare raises grave ethical problems for the institute or system that makes the decision and is responsible for it, as well as possible negative impacts

that can be devastating. In the sphere of healthcare care, the immutability feature of blockchain technology, despite the opportunities it opens to improve the integrity of the data, can suggest ethical questions. If an error is made in the medical records, the records are permanent on the blockchain, and this can be a disadvantage because there can be no corrections once the medical record is recorded this way. For example, a blockchain can lock information in permanent records. Although there is transparent and secure management of patient data via blockchain, this fixed schema creates a conflict between the security management of the patient database and the need to provide up-to-date patient records. Although machine learning and blockchain can increase security, they are not impermeable to security threats.

Most AI algorithms learn from the data fed to the system and therefore reproduce bias. For example, algorithms that act in healthcare facilities might inefficiently diagnose the level of some demographic groups because the training data samples contain insufficient diverse information, which can lead to ethical issues such as fairness and equality. W. Jung et al., (2023), healthcare systems have also used outdated legacy software that is still in use owing to its critical functionality but is often no longer supported by manufacturers, making them vulnerable to security risks and compatibility issues Affia et al. (2023).

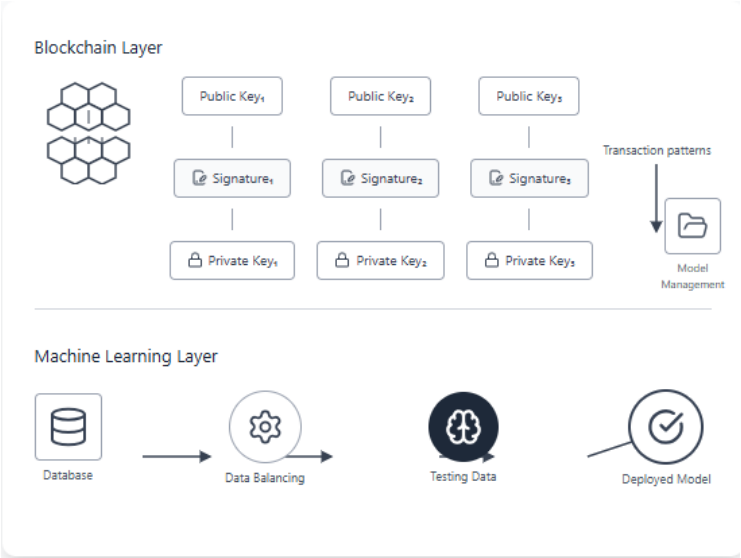


Figure 4: Challenges in Blockchain and Machine Learning

6. Advancing Cyber-Physical Systems

The creation of universals of where CPISs should be deployed has resulted in the problem of interoperability and security, given that the application is implemented haphazardly in different sectors. Such standards would standardize these processes by establishing normative measures for data protection, secure information exchange, and system links. Some of these standards could reduce the level of difficulty associated with deployment and guarantee the level of standardization needed, hence enhancing the availability and reliability of CPIS worldwide.

The modularity of CPIS is sometimes restricted by the associated need for computational, energy, and infrastructure resources. Future research efforts should focus on algorithms with low energy requirements and the corresponding system structures, in addition to edge computing, to facilitate the large-scale implementation of B-MAC. Improved scalability will create a new CPIS that can address different application requirements on different scales and regions.

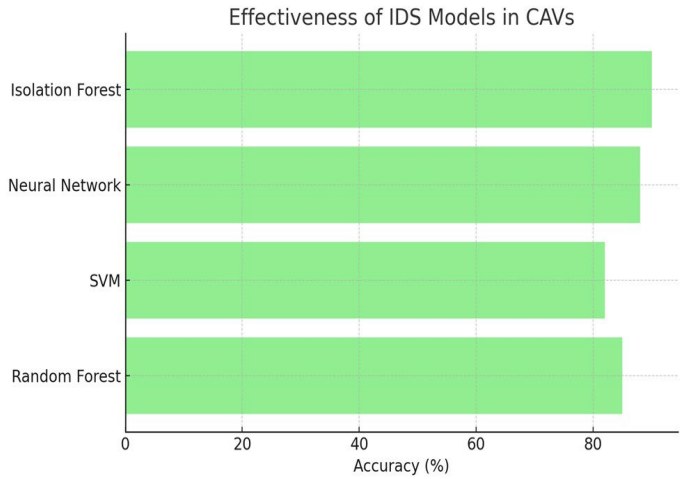


Figure 5: Effective of IDS Models in CAVs

7. Discussion

7.1 Connected and Automated Vehicles and Safety

Safety is the least we can expect from Connected and Automated Vehicles (CAVs), with many promising advances that come with performance improvements by the automotive industry and research findings, together with enhanced user experience developed through new services to help with traffic flow integration. (e.g., development of new information services) when CAVs become a part of everyday commuting. The "core" of road planning modules for a CAV system is algorithmically described within the system Ahmad et al. (2024).

CPIS is a form of bio resilience, and cybersecurity interface developed to identify recoverable common threats to cyber-physical threats in interconnected designs, which can mimic a biologically immutable system that can only be useful alone. infrastructures such as healthcare networks, smart cities, and connected automated vehicles (CAVs).

7.2 Policy Implication and Challenges

The importance of cybersecurity and road safety, as noted in previous sections, must be put into action by policymakers to increase the implementation of CAVs in traffic during medical emergencies. This study presents a summary of key cybersecurity problems in connected and automated vehicles (CAV) and their long-term impact on road safety, as well as the solutions proposed in the current literature. The significance of cyber-attacks and malicious information. The deployment of Connected and Automated Vehicles (CAVs) in traffic scenarios during medical emergencies requires policymakers to address essential cybersecurity and road safety issues. This study emphasizes several critical aspects of the problem.

8. Conclusion

The coupling of physical systems with AI can help CPIS leverage threat detection and mitigation, outperforming conventional methodologies to propose preventive measures against evolving threats. For the maximum effectiveness of the CPIS concept, its implementation requires interdisciplinary cooperation. We explain that AI and cyber security, engineers, policymakers, and social science professionals are critical and essential to join forces. Future studies propose the integration of better AI defense systems into existing CPIS models. As novel technologies continue to emerge in this field, they can shift from reactive to predictive security, allowing organizations to protect themselves from new threats in global markets. However, AI-based patient-oriented CPIS can provide real-time threat data learning, thus improving the use of ML algorithms for threat assessment. In the long run, these systems will develop the capacity to minimize the significant impact of cybercrimes and system failures.

AI and Ethics Declaration: This research was conducted in accordance with established academic ethical standards. Artificial intelligence tools were used solely to support language refinement formatting consistency and structural clarity. No AI systems were used to generate original research findings experimental results datasets or technical conclusions. All system designs methodologies simulations analyses and interpretations were conceived, developed and validated by the authors. The study did not involve human subjects' personal

data or clinical interventions and therefore did not require institutional ethics approval. The research focuses on cybersecurity risk analysis system modeling and controls experimental evaluation of connected cyber physical and medical technologies. All experiments were conducted in simulated or laboratory-controlled environments without impact on operational or patient-facing systems. The authors affirm full responsibility for the integrity, originality and accuracy of the work and confirm compliance with the ethical guidelines of the ICCWS ACI conference.

References

Ajakwe, S.O., Nwakanma, C.I., Kim, D. and Lee, J.M. 2022. Key wearable device technologies parameters for innovative healthcare delivery in B5G network. A review. IEEE Access, 10, pp. 49956 to 49974.

Alghamdi, T.A., Khalid, R. and Javaid, N. 2024. A survey of blockchain based systems. Scalability issues and solutions, applications and future challenges. IEEE Access, XX(YY), pp. xx to yy.

Alnasser, A., Sun, H. and Jiang, J. 2019. Cyber security challenges and solutions for V2X communications. A survey. Computer Networks, 151, pp. 52 to 67.

Asaju, B.J. 2024. Privacy preservation techniques in V2X ecosystems. Safeguarding individual privacy in connected vehicle environments. Journal of Artificial Intelligence Research, 4(1), pp. 58 to 72.

Arquilla, J., 2021. *Bitskrieg: the new challenge of cyberwarfare*. John Wiley & Sons.

Bajpai, A., Pandey, R., Singh, A.P., Jain, M. and Mishra, S. 2020. Article title placeholder. International Journal of Intelligent Communication and Computer Science, 1(1), pp. xx to yy.

Boubaker, S., Alsubaei, F.S., Said, Y. and Ahmed, H.E. 2023. Lightweight cryptography for connected vehicles communication security on edge devices. Electronics, 12(19), p. 4090.

Chaterji, S., Modi, A. and Singh, R. 2018. Impact of technology enablement on employee engagement. Corporate Communications: An International Journal, 25(1), pp. 132 to 153.

Chen, L., Zhang, C. and Ness, S. 2020. AI based intrusion detection for IoT. Obstacles, resolutions and prospective changes. Journal of Computer Science and Technology, 35(3), pp. 447 to 460.

Das, S., Siroky, G.P., Lee, S., Mehta, D. and Suri, R. 2021. Cybersecurity and the need for data and patient safety with cardiac implantable electronic devices. Heart Rhythm, 18(3), pp. 473 to 481.

Dorn, T.M. 2023. US critical infrastructure. Its importance and vulnerabilities to cyber and unmanned systems. Page Publishing Inc., New York.

Farooq, M.U. and Zhu, Q. 2020. Blockchain in emerging IoT applications. Review and open research challenges. Internet of Things, 11, p. 100240.

Ferrag, M.A., Derhab, A., Maglaras, L., Mukherjee, M. and Janicke, H. 2018. Privacy preserving schemes for fog based IoT applications. Threat models, solutions and challenges. In Proceedings of the International Conference on Smart Communications and Networking Technologies SaCoNeT. IEEE, New York, pp. 37 to 42.

Gatouillat, A., Badr, Y., Massot, B. and Sejdic, E. 2018. Internet of Medical Things. A review of recent contributions dealing with cyber physical systems in medicine. IEEE Internet of Things Journal, 5(5), pp. 3810 to 3822.

Hakeem, S.A.A., Abd El Gawad, M.A. and Kim, H. 2020. Comparative experiments of V2X security protocol based on hash chain cryptography. Sensors, 20(19), p. 5719.

Hasan, H.R., Salah, K. and Jayaraman, R. 2020. A blockchain based framework for secure and trustworthy smart manufacturing. Blockchain Research and Applications, 1(1 to 2), p. 100001.

Jung, W., et al. 2023. IoT health devices. Exploring security risks in the connected landscape. IoT, 4(2), pp. 150 to 182.

Liu, J., Zhang, W., Badr, J., Wang, L. and Qingjin. 2020. Cyber security challenges and solutions for V2X communications. A survey. Computer Networks, 151, pp. 52 to 67.

Morris, R., Trivedi, M.M. and Pan, S.J. 2009. Preferences for treatment outcomes among patients with panic disorder. IEEE Transactions on Wireless Communications, 58(5), pp. 1 to 14.

Pandey, R.K. and Misra, M. 2016. Cyber security threats in smart grid infrastructure. In National Power Systems Conference NPSC 2016. IEEE, New York, pp. 1 to 6.

Pantopoulou, L., Kritikos, K. and Magoutas, B. 2020. Machine learning in IoT security. Current trends and future directions. Future Generation Computer Systems, 112, pp. 1 to 13.

Polo, L. 2024. Revolutionizing sales and operations planning with artificial intelligence. Insights and results. International Journal for Multidisciplinary Research, 6(6), pp. xx to yy.

Rammohan, A., et al. 2023. Revolutionizing intelligent transportation systems with cellular vehicles to everything technology. Current trends, use cases, emerging technologies, standardization bodies, industry analytics and future directions. Vehicular Communications, 43, p. 100638.

Rao, P.M., Jangirala, S., Pedada, S., Das, A.K. and Park, Y. 2023. Blockchain integration for IoT enabled V2X communications. A comprehensive survey, security issues and challenges. IEEE Access, 11, pp. 54476 to 54494.

Yogeswaran, M. and Ponnambalam, S.G. 2010. Swarm robotics. An extensive research review. In Advanced Knowledge Application in Practice. InTech, Rijeka, pp. 259 to 278.

Yu, W., Liang, F., He, X., Hatcher, W.G., Lu, C., Lin, J. and Yang, X. 2017. A survey on edge computing for the Internet of Things. IEEE Access, 6, pp. 6900 to 6919.

Zhang, X., Jiang, X. and Yang, X. 2019. Blockchain for IoT. A survey and research directions. Future Generation Computer Systems, 92, pp. 268 to 283.

The Evolution of Penetration Testing in the Era of AI

Errol Baloyi, Mpho Letshwenyo, Mamello Mtshali and Alex Ramantswana
Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa

EBaloyi2@csir.co.za
MLetshwenyo@csir.co.za
MMtshali3@csir.co.za
ARamantswana@csir.co.za

Abstract: Over the past several decades, penetration testing has transitioned from a predominantly manual, expert-driven activity to a mature discipline supported by automation, modular frameworks, and artificial intelligence (AI)-assisted tools. This study provides a descriptive review of the historical evolution of penetration testing tools, highlighting the major technological and methodological advancements that have shaped the field. In addition, a practical comparative evaluation of two widely used tools, Burp Suite Professional and the Open Worldwide Application Security Project (OWASP) Zed Attack Proxy (ZAP) was conducted using a controlled vulnerable web application, Damn Vulnerable Web Application (DVWA), to assess their performance and usability in a realistic testing environment. The study further examines the impact of AI on the contemporary and emerging landscape of penetration testing tools. The findings suggest that AI is augmenting existing tools through enhanced automation and more effective vulnerability identification, while simultaneously enabling new paradigms in both offensive and defensive cybersecurity practices. This work contributes to the understanding of the evolving role of penetration testing in an AI-influenced context and discusses the implications of these developments for researchers, practitioners, and tool developers.

Keywords: Artificial intelligence, Penetration testing, Cybersecurity, Vulnerability assessment

1. Introduction

Today, amid the digital era, cyberspace is constantly under threat of being compromised by malicious actors. Often referred to as cyber threat actors (CTAs), these individuals or groups intentionally cause harm or engage in malicious activities within cyberspace. However, not all threat actors are the same, they differ in terms of motivation, resources, and skill. This distinction is illustrated by Sailio, Latvala and Szanto (2020), who categorize threat actors as nation-state sponsored, cybercriminals, ideologically driven attackers, insider threats, or thrill seekers. According to Ravindran and Potukuchi (2022), the number of threat actors around the world is likely to continue growing due to the increasing number of web applications. Furthermore, the exponential growth of the Internet has worsened the situation, with the emergence of large language models (LLMs) which have further complicated the cyber landscape.

The arrival of new and advanced applications, while rich in features, has also increased system vulnerabilities. Which according to Deng et al (2024), securing these systems has become a formidable challenge. As a result, offensive security methods such as vulnerability assessment, penetration testing (VAPT) have become essential components of the cybersecurity lifecycle. To fully grasp the concept of VAPT, it is essential to first understand what constitutes a vulnerability in a system. According to Khera et al (2019), a vulnerability is a flaw or loophole in an application that allows a threat actor to gain unauthorized system privileges and access sensitive data, which can then be exploited for malicious purposes. Similarly, Goel and Mehtre (2015) define a vulnerability as a weakness, either due to a bug in implementation or a design flaw that enables an attacker to harm users and escalate their access privileges.

1.1 Vulnerability Assessment vs. Penetration Testing

Today, many institutions often find themselves in situations where they are required whether by law, stakeholders, or customers to secure their systems and data against a growing number of vulnerabilities. This has made the use of VAPT increasingly necessary. According to Khera et al (2019), VAPT helps organizations proactively discover and identify vulnerabilities before attackers can exploit them. Rahman (2025) further emphasises that VAPT enables organizations to detect security loopholes that could otherwise be leveraged by malicious actors to launch attacks against their systems.

Thus far, in the current study, vulnerability assessment and penetration testing have been used interchangeably, however, they differ in scope and purpose. According to Khera et al (2019), vulnerability assessment is the process of identifying and discovering weaknesses or loopholes in a system using automated tools such as Nessus. These vulnerabilities can provide potential entry points for attackers. In contrast, penetration testing (or pentesting) is typically a manual process that follows a vulnerability assessment and involves actively

exploiting the identified weaknesses in a controlled and systematic manner usually after obtaining proper authorization from the organization. Nevertheless, for the purposes of this study, the term "penetration testing" is used to collectively refer to both vulnerability assessment and penetration testing.

1.2 The Importance of Penetration Testing in Cybersecurity

Penetration tests are conducted to identify vulnerabilities within a system and to determine how they can be mitigated. These tests involve simulating various types of attacks on the target system, effectively mimicking the behaviour of a malicious actor (Alhamed and Rahman, 2023). Vulnerabilities across different information assets are assessed either manually or through automated tools (Moreno et al, 2025). It is important to clarify that a penetration test is performed by a penetration tester (pentester), not a threat actor (Božić, Penevski and Adamović, 2019). The pentester is typically tasked with identifying security flaws in areas such as web applications, network configurations, or the overall infrastructure design of the organization that has engaged their services. According to Khera et al (2019), many organizations suffer from poorly designed software and hardware, as well as system misconfigurations.

As a result, penetration testing is a critical practice not only does it help address these specific issues, but it also acknowledges the reality that achieving a completely vulnerability-free system is nearly impossible. However, by identifying and eliminating as many vulnerabilities as possible through penetration testing, organizations can significantly strengthen their security posture (Khera et al, 2019; Goel and Mehtre, 2015). For instance, the 2013 Adobe security breach, which exposed data of approximately 153 million users (BBC, 2025). The leaked data included user IDs, names, and encrypted information such as passwords and credit card details. In response, Adobe issued a white paper in which it assured clients of its commitment to data security. The company emphasized that it was conducting regular penetration testing of its products and services to identify and fortify weak points (Adobe, 2024). Since then, Brown (2025) notes that this ongoing testing strategy has helped Adobe consistently reinforce its defences, preventing any further major breaches of comparable scale.

2. Research Methodology

This paper adopts a descriptive review approach to analyse the evolution of penetration testing tools from before the year 2000 through to the present. Tools were selected based on their historical significance, widespread usage, and representation of emerging trends in penetration testing, especially as influenced by AI in recent years. Each tool was assessed using a consistent framework, which included the following criteria: creator, description, primary use, and licensing. Sources included official tool documentation, peer-reviewed publications, cybersecurity blogs, community forums, and public repositories such as GitHub. The review spans more than two decades to highlight shifts in design philosophy, usage patterns, and the growing integration of AI in modern tools.

To supplement the review, a basic practical assessment was conducted using a controlled testbed. Burp Suite Professional and OWASP ZAP were selected for the hands-on comparison due to their prominence in web application security testing and their contrasting licensing models (commercial versus open source). Furthermore, Burp Suite Professional incorporates vendor-supported AI integration to enhance vulnerability analysis and automation, whereas OWASP ZAP lacks native AI functionality and instead depends on third-party or community-developed extensions to achieve comparable capabilities. The widely used DVWA platform was employed as the target environment, chosen for its reproducible vulnerabilities and compatibility with the scanners.

3. The Genesis and Evolution of Penetration Testing

The term hacking originated in the 1960s at the Massachusetts Institute of Technology (MIT), specifically within the Tech Model Railroad Club (TMRC), where members sought to "hack" circuits to enhance the performance of their model trains (Whitaker and Newman, 2006). Over time, the term evolved to describe activities involving unauthorized access to computer systems. Despite this shift in association, the initial push for the development of advanced computer penetration techniques came from the United States Department of Defence (Hunt, 2012).

At that time, computers were not yet widespread, and the concept of pentesting was largely confined to military and intelligence communities, where it was known as “red teaming” a practice involving controlled, adversarial simulations to evaluate defensive capabilities (Horkan, 2025). A significant turning point occurred in the late 1960s during the Joint Computer Conference, where the need to assess digital systems for security vulnerabilities was formally recognized within both academic and operational contexts. This period also

coincided with the military's use of national time-sharing computer systems, further reinforcing the importance of such security evaluations (Hunt, 2012).

In 1969, Advanced Research Projects Agency Network (ARPANET), a United States defence program was launched as a precursor to the modern Internet, funded by the Advanced Research Projects Agency (ARPA), later renamed Defence Advanced Research Projects Agency (DARPA). ARPANET fostered a culture of exploration, which contributed to the emergence of early hacking practices (DARPA, 2025). This period also saw the formation of "Tiger Teams," elite units within the National Security Agency (NSA) and the Department of Defence, tasked with attempting to breach United States military computer systems (Horkan, 2025). These efforts are widely regarded as the first instances of penetration testing, with the Tiger Teams recognised as the earliest practitioners of the discipline (Palatty, 2025; Hunt, 2012; Schaefer, 2004).

Furthermore, these teams served as early prototypes for the red team/blue team model that is now widely employed in modern cybersecurity exercises (Palatty, 2025). However, as the field was still in its infancy, the "Tiger Teams" experienced limited success. This prompted the development of the 1972 report by J.P. Anderson, which provided detailed step-by-step guidelines to enhance the effectiveness of such teams. Remarkably, many of the procedures outlined in Anderson's 1972 report continue to form the foundational framework for contemporary penetration testing practices (Palatty, 2025; Hunt, 2012; Schaefer, 2004).

3.1 Early Days: Manual Exploitation and Custom Scripts

The expansion of personal computing and enterprise networks during the 1980s and 1990s corresponded with a growing need for systematic approaches to security testing. Penetration testing, once confined to classified government operations, increasingly entered the private sector in response to the emerging vulnerabilities associated with networked systems and the early Internet (Horkan, 2025). The introduction of some of the penetration testing tools discussed below marked a critical development in the evolution of the practice.

Table 1: Pre-2000s Tools

Tool	Creator	Description	Primary Use	Licensing
Internet Security Systems Scanner (ISS Scanner)	Developed by Christopher Klaus in 1992 at the Georgia Institute of Technology and acquired by IBM in 2006.	A pioneering tool designed to identify security vulnerabilities within networked systems.	Network testing, providing administrators insights into potential weaknesses.	Commercial; integrated into IBM's security offerings. References: (Cummings, 2025; NBC News, 2006).
Security Administrator Tool for Analysing Networks (SATAN)	Developed by Dan Farmer and Wietse Venema, and released on April 5, 1995, it was later superseded by the MITRE CVE system.	A software vulnerability scanner with web interface and context-sensitive tutorials, using signatures to identify common network-related problems.	Automated identification of known network-exploitable vulnerabilities.	Open source. References: (Network Encyclopaedia, 2024).
Ethereal/Wireshark	Ethereal was developed by Gerald Combs in 1997 and renamed Wireshark in 2006.	Network protocol analyser with deep packet inspection capabilities for real-time traffic capture and interception.	Analysing virtual network traffic in VPCs; identifying malicious activities and abnormal network behaviours.	Licensed under the GNU General Public License, version 2. References: (Anand and Singh, 2021; Nadimpalli et al, 2025; O'Neill, 2025).
Nmap or Network Mapper	Developed by Gordon "Fyodor" Lyon in 1997.	Is a tool used to scan networks to find hosts that are running on a network, identify the services within that network such as a web server and discover security vulnerabilities	Network discovery and security auditing.	Open source. References: (Adam, Widyawan and Putra, 2023; Lyon, 2025; Singh et al, 2024).

Tool	Creator	Description	Primary Use	Licensing
Snort	Developed by Martin Roesch, founder of Sourcefire in 1998, and acquired by Cisco in 2013.	A network intrusion detection and prevention system (IDPS) with real-time traffic analysis, packet logging, and detection capabilities.	Network testing (intrusion detection/prevention); detecting probes, attacks, buffer overflows, and port scans.	Licensed under the GNU General Public License (GPLv2+) and actively maintained by Cisco. References: (Lenaerts-Bergmans, 2025).

3.2 The Rise of Frameworks and Automation (2000s - Early 2010s)

The early 2000s introduced a more systematic and standards-driven era of penetration testing. The creation of the Open Web Application Security Project (OWASP) in 2001 played a pivotal role in this transition, offering guidance for testing web applications, specifically with the release of OWASP first penetration testing framework in 2003 (Palatty, 2025). Consequently, the current study details below the pen-testing tools that emerged during that period.

Table 2: 2000s-Early 2010s Tools

Tool	Creator	Description	Primary Use	Licensing
Open-Source Security Testing Methodology Manual (OSSTM)	Developed by Pete Herzog in 2001 at the Institute for Security and Open Methodologies.	It is a comprehensive, peer-reviewed methodology for operational security assessments covering five core channels.	Vulnerability and security assessments, penetration testing, ethical hacking.	Open source. References: (Raynaud, 2025).
Nessus	Developed by Renaud Deraison in 1998 and commercialized by Tenable Network Security in 2002.	An automated vulnerability scanning tool that detect security flaws in networks, servers, applications, and cloud environments.	Penetration testing, compliance auditing (PCI DSS, HIPAA, CIS), continuous security monitoring.	Commercial and open-source alternative is OpenVAS. References: (Deraison, 2025; Raynaud, 2025).
Metasploit Framework	Developed by H.D. Moore in 2003 and acquired by Rapid7 in 2009.	A penetration testing framework with extensive database of exploits, payloads, and post-exploitation modules.	Network testing, web application testing, exploitation, post-exploitation, security validation.	Primarily open source (BSD license); commercial editions: Metasploit Express and Pro. References: (Chipeta, 2025; Kennedy et al, 2011; Nadimpalli et al, 2025).
Burp Suite	Developed by Dafydd Stuttard in 2003 and later evolved by PortSwigger, Dafydd Stuttard's security company.	A comprehensive web application security testing suite acting as intercepting proxy with scanning, spidering, and attack capabilities.	Web application penetration testing, and vulnerability assessments.	Proprietary; free Community Edition and paid Professional/Enterprise editions. References: (PortSwigger, 2025).
Acunetix	Developed by Nick Galea in 2005.	Specialized web application security testing tool with automated scanning for SQL Injection, XSS, and other vulnerabilities.	Web application testing and auditing.	Commercial. References: (Acunetix, 2025; Briffa, 2025).
Aircrack-ng	Developed by Thomas 'Mister X' D'Otreppe in 2006.	It is a collection of tools for assessing Wi-Fi network security, monitoring, and cracking encryption.	Wireless network security testing and assessment.	Open source. References: (Cooper, 2025; Hurley et al, 2007).

3.3 Specialization and Cloud Integration (Mid-2010s - Present)

Since the mid-2010s, penetration testing has undergone a pivotal transformation driven by cloud adoption, DevOps practices, and increasingly sophisticated cyber threats fuelled by growing Internet usage. This shift emphasized established security frameworks such as the Payment Card Industry Data Security Standard (PCI DSS), to become increasingly critical amid rising online fraud (Fitzgerald and Rubbinaccio, 2025; Barney, 2025). Tool development evolved along three paths: legacy tools adapted for cloud use, cloud-native tools addressing risks like Infrastructure as Code (IaC) misconfigurations, and evasion techniques targeting cloud-native defences (Morris, 2025).

Penetration testing frameworks matured to guide methodologies and tool selection, with the OWASP Testing Guide endorsing tools like Burp Suite for API testing and WPScan for CMS vulnerabilities. Standards such as PTES structured assessments through phases like threat modelling and post-exploitation, leveraging tools like Nmap and Metasploit, while the NIST Cybersecurity Framework standardized compliance with tools like Nessus (Adam, Widyawan and Putra, 2023). As cloud environments grow in complexity, tools continue to specialize, spanning cloud-native scanners, cloud-adapted legacy solutions, payload generators, and advanced evasion tools (Mallick and Nath, 2024).

Table 3: Mid-2010s - Present Tools

Tool	Creator	Description	Primary Use	Licensing
Scout Suite	Developed by NCC group.	It is a multi-cloud security auditing tool using APIs to collect configuration data and generate HTML reports categorizing findings by severity.	Security posture assessments across AWS, Google Cloud, and Microsoft Azure; investigating security features across cloud services.	Licensed under the GNU General Public License v2 (GPLv2). References: (Hossam, 2025; Kesseli, 2025).
WPScan	Developed by the WPScan Team and now owned by Automattic Inc.	Vulnerability assessment tool for WordPress with highest GitHub metrics.	WordPress-specific security assessments including plugins, themes, and error log exposure.	Licensed under a dual-license model called the WPScan Source License. References: (Farrell, 2022).
Phantom Evasion	Developed by oddcod3 and archived in a GitHub repository.	Antivirus evasion tool generating fully undetectable payloads via obfuscation and reducing detection.	Payload generation for cloud penetration testing; limited by cloud behavioural analysis and runtime protection.	Licensed under the GNU General Public License version 3 (GPL3). References: (Adam and Sufyanu, 2021; Tigner, Wimmer and Rebman, 2021).
MSFVenom	Part of Metasploit Framework	Payload generation tool creating cloud-optimized payloads that evade defences through encrypted reverse TCP sessions.	Bypassing sandboxed cloud VMs.	Part of Metasploit (BSD license). References: (Pandey, 2025; Tigner, Wimmer and Rebman, 2021).

4. Penetration Testing Tools in the Age of AI

The current study has so far provided a detailed examination of traditional penetration testing tools. While these methods have been effective in the past, they often struggle to keep up with today's rapidly evolving threat landscape as also noted by (Thummala, Gupta and Siddharth, 2025). In response to this challenge, automated penetration testing has emerged as a viable solution, enabling quicker and more consistent scanning and testing of networks, applications, and systems for vulnerabilities. Accordingly, the following section explores the implications of pentesting tools in the era of AI. However, it is important to note that the tools mentioned below do not represent the full range of tools available. As Goel and Mehtre (2015) points out, there are numerous open-source and commercial pentesting tools currently on the market. These tools are developed by industry experts with the aim of delivering more efficient, dynamic, and comprehensive security assessments (Božić, Penevski and Adamović, 2019).

4.1 AI as an Augmentation of Existing Tools

AI in penetration testing has changed traditional methods of testing by improving existing tools through methods like automated vulnerability prioritization, adaptive threat simulation, and predictive analytics. This innovation helps in fixing issues that exist in traditional methods like high false positives, scalability issues in complex environments and inconsistent workflows. AI-powered tools like Tenable.io prioritize vulnerabilities by linking threat intelligence exploit feasibility, and asset criticality; this in turn reduces manual triage efforts (Hossam, 2025). Integrated tools like AIPenPool combine legacy tools like Nmap and Metasploit with Machine Learning (ML) algorithms, this improves detection accuracy while reducing false positives by using verification algorithms that compare vulnerability data against historical records (Kumar et al, 2025).

AI improves offensive capabilities by simulating advanced persistent threats (APTs). Tools like Microsoft Security Copilot mimic complex attack patterns by learning from past breaches. This allows for dynamic red-team exercises that respond to defensive actions in real-time (Hossam, 2025). Platforms like Cyberwheel automatically run attack chains that follow the MITRE ATT&CK framework by using reinforcement learning (RL). They choose methods based on the network's weaknesses and its current state. This is an improvement from scripted tools like MITRE Caldera, which uses various hacking tools but does not have the ability to make decisions on their own (Kumar et al, 2025). RL agents and Cyberwheel can apply their capabilities in different situations, they maintain strong performance even with new network setups, something traditional automated tools struggle to do.

Operational efficiency is also improved through AI integration, automated workflows seamless chain tools, reducing manual intervention. This makes security accessible to non-experts through user-friendly interfaces, it also increases the capability of organizations without sacrificing depth. For instance, AIPenTool's guided attack suggestions enable users to execute complex tests like SQL injection or SSL vulnerability assessments with minimal expertise (Kumar et al, 2025). Despite these improvements, challenges still exist. Due to noisy inputs reducing the accuracy of ML models, data quality is crucial. AI tools are still not widely used in the real world. Simulators like Cyberwheel have not yet gone through extensive testing in real-world settings. Additionally, there are some ethical concerns that have been raised, requiring frameworks like Safe Reinforcement Learning to prevent uncontrolled autonomous attacks. Semi-autonomous systems allow AI to suggest attack paths while needing human approval, balancing autonomy with ethical oversight (Fernandes et al, 2025). Additionally, integrating AI with older tools requires a lot of changes, especially for real-time adaptation to unknown vulnerabilities. However, below the current study provides a comparison of an AI integration versus non-AI approaches in automated web application testing tool.

4.2 Practical Demonstration: AI vs Non-AI Tool

As outlined in the methodology, the present study conducts a comparative analysis of two web application testing tools: Burp Suite Professional, a widely used commercial solution, and OWASP ZAP 2.16.1, an open-source counterpart. A summary of the comparative results is presented in Figure 1.

BURP

Severity	Confidence: Certain	Firm	Tentative	Total
High	5	1	0	6
Medium	4	7	1	12
Low	1	0	0	1
Information	36	7	0	43
False Positive	0	0	0	0

ZAP

Risk	Confidence: User Confirmed	High	Medium	Low	Total
High	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Medium	0 (0.0%)	2 (18.2%)	2 (18.2%)	0 (0.0%)	4 (36.4%)
Low	0 (0.0%)	1 (9.1%)	3 (27.3%)	0 (0.0%)	4 (36.4%)
Informational	0 (0.0%)	1 (9.1%)	2 (18.2%)	0 (0.0%)	3 (27.3%)
Total	0 (0.0%)	4 (36.4%)	7 (63.6%)	0 (0.0%)	11 (100%)

Figure 1: VA Scan Results

The results presented in Figure 1 are derived from basic automated scans. Both tools demonstrated strong performance. However, Burp exhibited superior detection capabilities compared to ZAP, which only identified 11 vulnerabilities, whereas Burp Suite detected a total of 61. Similar findings were reported by Albalawi et al (2023), who observed that ZAP generally reports fewer vulnerabilities than Burp. Nonetheless, Albalawi et al

(2023) also noted that Burp scans typically take longer, a pattern the present study also observed. Similarly, Deeb (2024) also employed DVWA as a testbed to evaluate several leading web application security tools, including ZAP and Burp Suite, and concluded that Burp Suite achieves an excellent balance between detection accuracy and resource utilization.

4.3 The Dual-Edged Sword: AI for Attackers

AI or ML, while critical in strengthening cybersecurity defences, also pose a pernicious threat by empowering threat actors to conduct increasingly sophisticated, AI-generated exploits (Lewis et al, 2025). According to Stanham (2025), threat actors are leveraging several key strategies to accelerate and enhance their attacks. One such strategy is the use of automated and adaptive social engineering, where AI is employed to craft convincing phishing emails, create fake websites, generate deepfake audio and visuals, inject malicious prompts or code, and carry out chatbot-driven fraudulent activities. Another method involves the evasion of security mechanisms.

Stanham (2025) further notes that attackers are using AI to bypass CAPTCHAs, poison security data models, and exploit malicious generative pre-trained transformers (GPTs) to undermine traditional defence systems. Additionally, AI-assisted vulnerability discovery, as highlighted by Kerner (2025), enables the rapid identification of software and system flaws through the analysis of large volumes of data, significantly accelerating the exploitation process. Finally, AI continues to enable threat agents to redefine and adapt their strategies in real time, posing an escalating challenge to existing cybersecurity frameworks.

5. Conclusion and Future Work

The findings of this study demonstrate that the process of conducting VAPT has evolved substantially since the early stages of information security and the tools employed for VAPT have also advanced in parallel. Previously, such practices were confined to the military and defence organizations due to the sensitive nature of their operations. However, with the expansion and increasing sophistication of the threat landscape, cybersecurity has become a fundamental concern across a wide spectrum of organizations. Moreover, the integration of AI has further transformed VAPT tools. This evolution is evident in the comparative assessment conducted in this study, where Burp, a tool incorporating AI-driven capabilities, identified a greater number of vulnerabilities compared to ZAP. Consequently, future research should prioritize the development of ethically aligned AI systems for cybersecurity, alongside the creation of frameworks that ensure accountability and transparency in AI-assisted penetration testing. Finally, future work will expand this study by incorporating additional web security assessment tools and benchmarking them against standardized evaluation metrics to enable a more comprehensive comparative analysis.

Ethics declaration: Ethical clearance was not required, as this study did not involve human participants.

AI declaration: AI was not utilised in this study.

References

Acunetix. (2025) “Acunetix - Crunchbase Company Profile & Funding”, [online], https://www.crunchbase.com/organization/acunetix.

Adam, A.S., and Sufyanu, Z. (2021) “Performance comparison of PyRAT and Phantom antivirus software evasion tools”, *Journal of Information Security*, Vol. 2, No. 1.

Adam, H.M., Widyawan and Putra, G.D. (2023) “A review of penetration testing frameworks, tools, and application areas”, *IEEE 7th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE 2023),* pp 319-324.

Adobe. (2024) “Adobe Application Security Overview”.

Albalawi, N., Alamrani, N., Aloufi, R., Albalawi, M., Aljaedi, A., & Alharbi, A. R. (2023) “The Reality of Internet Infrastructure and Services Defacement: A Second Look at Characterizing Web-Based Vulnerabilities”, *Electronics*, Vol. 12, No. 12.

Alhamed, M. and Rahman, M.M.H. (2023) “A Systematic Literature Review on Penetration Testing in Networks: Future Research Directions”, *Applied Sciences*, Vol. 13, No. 12, p 6986.

Anand, P. and Singh, A.S. (2021) “Penetration testing security tools: A comparison” *2021 10th International Conference on System Modeling and Advancement in Research Trends (SMART 2021),* pp 182–184.

Barney, N. (2025) “What is PCI DSS? Requirements and Compliance”, [online], https://www.techtarget.com/searchsecurity/definition/PCI-DSS-Payment-Card-Industry-Data-Security-Standard.

BBC. (2025) “Adobe hack: At least 38 million accounts breached”, [online], https://www.bbc.com/news/technology-24740873.

Božić, K., Penevski, N. and Adamović, S. (2019) “Penetration Testing and Vulnerability Assessment: Introduction, Phases, Tools and Methods”, *SINTEZA 2019*, pp 229-234.

Briffa, M. (2025) “Maltese Success Story: Acunetix World Leading in Cybersecurity” [online], https://newtech.mt/blog-news/maltese-success-story-acunetix-world-leading-in-cybersecurity/.
Brown, C. (2025) “Penetration Testing Examples: Real-World Scenarios & Insights”, [online], https://www.vikingcloud.com/blog/penetration-testing-examples.
Chipeta, C. (2025) “What is Metasploit?”, [online], https://www.upguard.com/blog/metasploit.
Cooper, S. (2025) “Aircrack-ng Review for 2024 & the Best Alternatives (Paid & Free)”, [online], https://www.comparitech.com/net-admin/aircrack-ng-review/.
Cummings, D. (2025) “Atlanta History: Notes from the ISS S-1 IPO Filing”, [online], https://davidcummings.org/2012/09/07/atlanta-history-notes-from-the-iss-s-1-ipo-filing/.
DARPA. (2025) “ARPANET”, [online], https://www.darpa.mil/news/features/arpanet.
Deeb R. (2024) “Evaluating Web Application Vulnerability Scanners: Introducing the RD-Score for Comprehensive Performance Assessment”, *International Journal of Open Information Technologies,* Vol. 12, No. 11.
Deng, G. et al. (2024) “PentestGPt: Evaluating and Harnessing Large Language Models for Automated Penetration Testing”, *USENIX Security Symposium 2024*.
Deraison, R. (2025) “Nessus Turns 20!”, [online], https://www.tenable.com/blog/nessus-turns-20.
Farrell, S. (2022) “Abstraction and automation of WordPress vulnerability scanning”, Unpublished manuscript.
Fernandes, R., Lopes, N., Goncalves, J., and Cosgrove, J. (2025) “Comparing traditional hacking tools and AI-driven alternatives”, *13th International Symposium on Digital Forensics and Security (ISDFS 2025).*
Fitzgerald, A. and Rubbinaccio, M. (2025) “PCI DSS History: How the Standard Came To Be”, [online], https://secureframe.com/blog/pci-history.
Goel, J.N. and Mehtre, B.M. (2015) “Vulnerability Assessment & Penetration Testing as a Cyber Defence Technology”, *Procedia Computer Science*, pp 710-715.
Horkan, W. (2025) “A Brief History of Penetration Testing: From Tiger Teams to PTaaS – Horkan”, [online], https://horkan.com/2025/06/19/a-brief-history-of-penetration-testing-from-tiger-teams-to-ptaas.
Hossam, A. (2025) “Best tools for cloud penetration testing in 2025”, [Online], https://deepstrike.io/blog/best-tools-for-cloud-penetration-testing-in-2025.
Hunt, E. (2012) “US government computer penetration programs and the implications for cyberwar”, *IEEE Annals of the History of Computing*, Vol. 34, No. 3, pp 4-21.
Hurley, C., Rogers, R., Thornton, F., Connelly, D. and Baker, B. (2007) “WarDriving and Penetration Testing with Windows”, *Syngress*, pp 93-117.
Infosec. (2025) “The history of penetration testing”, [online], https://www.infosecinstitute.com/resources/penetration-testing/the-history-of-penetration-testing/.
Kennedy, D., O’Gorman, J., Kearns, D. and Aharoni, M. (2011) “Metasploit: The Penetration Tester’s Guide”, No Starch Press.
Kerner, S.M. (2025) “AI-powered attacks: What CISOs need to know now”, [online], https://www.techtarget.com/searchsecurity/feature/AI-powered-attacks-What-CISOSs-need-to-know-now.
Kesseli, T. (2025) “Security in cloud environments”, Unpublished manuscript.
Khera, Y., Kumar, D., Sujay, S. and Garg, N. (2019) “Analysis and Impact of Vulnerability Assessment and Penetration Testing”, *International Conference on Machine Learning, Big Data, Cloud and Parallel Computing*, pp 525-530.
Kumar, N.P., Vijay Kumar Gowda, K.K., and Dhanushree, M.S. (2025) “AIPenTool: A unified approach to automated penetration testing for enhanced network and web application security”, *3rd International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE 2025)*.
Lenaerts-Bergmans, B. (2025) “Snort Explained: Understanding Snort Rules and Use Cases”, [online], https://www.crowdstrike.com/cybersecurity-101/threat-intelligence/snort-rules/.
Lewis, C., Kristensen, I., Caso, J. and Fuchs, J. (2025) “AI is the greatest threat—and defense—in cybersecurity today. Here’s why”, [online], https://www.mckinsey.com/about-us/new-at-mckinsey-blog/ai-is-the-greatest-threat-and-defense-in-cybersecurity-today.
Lyon, G. (2025) “Nmap Preface. Nmap Network Scanning”, [online], https://nmap.org/book/preface.html.
Mallick, A.I. and Nath, R. (2024) “Navigating the cybersecurity landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments”, [Online], https://worldscientificnews.com/navigating-the-cyber-security-landscape-a-comprehensive-review-of-cyber-attacks-emerging-trends-and-recent-developments/.
Moreno, A.C. et al. (2025) “Analysis of Autonomous Penetration Testing Through Reinforcement Learning and Recommender Systems”, *Sensors*, Vol. 25, No. 1.
Morris, K. (2025) “Infrastructure as Code: Designing and Delivering Dynamic Systems for the Cloud Age”, O’Reilly Media, Sebastopol.
Nadimpalli, A.V., Sai, A.S. and Kamalesh, M.D. (2025) “Analyzing and implementing cloud security and penetration testing tools: A review”, *Journal of Cloud Security*, Vol. 8, No. 2, pp 110-125.
NBC News. (2006) “IBM to acquire Internet Security Systems”, [online], https://www.nbcnews.com/id/wbna14482359.
Network Encyclopaedia. (2024) “Satan tool: Security administrator tool for analyzing networks. NETWORK ENCYCLOPEDIA”, [online], https://networkencyclopedia.com/security-administrator-tool-for-analyzing-networks-satan/.
O’Neill, T. (2025) “Wireshark – from ethereal to today. Network Visibility”, [online], https://www.garlandtechnology.com/blog/wireshark-from-ethereal-to-today.

Palatty, N.J. (2025) “A Brief History of Penetration Testing - Astra Security Blog. Astra Security Audit”, [online], https://www.getastra.com/blog/security-audit/history-penetration-testing/?utm_source=chatgpt.com.

Pandey, B., Shukla, M.M., Pandey, P., and Bakar, W.A. (2025) “Development of malware in Windows sandbox and Kali MSFvenom for hacking Android and Windows operating systems”, *International Journal of Information Technology*, Vol. 17, No. 3, pp 1831–1838.

Pautov, A. (2024) “OWASP ZAP: A comprehensive guide to web application security testing, Medium”, [online], https://medium.com/@1200km/owasp-zap-a-comprehensive-guide-to-web-application-security-testing-6c247f4be39b.

PortSwigger. (2025) “Burp Suite documentation”, [online], https://portswigger.net/burp/documentation.

Rahman, H. (2025) “Real-Time Network Traffic Monitoring System”, [online], https://scde.jntuh.ac.in/lmsportal/alumniuploads/programdocuments_29052025/Network_monitoring_system_Using_wireshark.pdf.

Ravindran, U. and Potukuchi, R.V. (2022) “A Review on Web Application Vulnerability Assessment and Penetration Testing”, *Review of Computer Engineering Studies*, Vol. 9, No. 1, pp 1-22.

Raynaud, C. (2025) “Open-Source Security Testing Methodology Manual”, [online], https://raynaudc.wordpress.com/2011/12/20/open-source-security-testing-methodology-manual-2/.

Sailio, M., Latvala, O.M. and Szanto, A. (2020) “Cyber Threat Actors for the Factory of the Future”, *Applied Sciences*, Vol. 10, No. 12, p 4334.

Schaefer, M. (2004) “If A1 is the answer, what was the question? An edgy naïf’s retrospective on promulgating the trusted computer systems evaluation criteria”, *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, pp 204-228.

Singh, A., Sharma, E.S., Reddy, B.K., Soni, P., Ghuman, S.S., and Gill, U.S. (2024) “Automated network vulnerability assessment with Nmap: A comprehensive approach*”, 2nd International Conference on Advanced Computing and Communication Technologies (ICACCTech 2024)*, pp 208–214.

Stanham, L. (2025) “Most Common AI-Powered Cyberattacks. CrowdStrike”, [online], https://www.crowdstrike.com/en-us/cybersecurity-101/cyberattacks/ai-powered-cyberattacks/.

Thummala, V.R., Gupta, A.K. and Siddharth, E. (2025) “Enhancing Cybersecurity with AI-Powered Penetration Testing Tools”, *1st Int. Conf. on Holistic Approaches to Achieve Sustainable Development Goals with AI (ICASDGAI-2025).*

Tigner, M., Wimmer, H. and Rebman, C.M. (2021) “Analysis of Kali Linux penetration tools: A survey of hacking tools”, *International Conference on Electrical, Computer, and Energy Technologies (ICECET 2021).*

Whitaker, A. and Newman, D.P. (2006) *Penetration Testing and Network Defense*, Cisco Press.

Bridging the Cyber Gap: Mapping Misalignment Between Digital Adoption and Cybersecurity Capacity

Zakariya Belkhamza
Ahmed Bin Mohammed Military College, Doha, Qatar

zbelkhamza@abmmc.edu.qa

Abstract: The rapid pace of digital transformation has revealed a structural asymmetry between technological expansion and protective capabilities. As digital adoption speeds up across economies, cybersecurity frameworks tend to develop more slowly, creating what this paper calls *the cyber gap*: a persistent misalignment between the trajectories of digital adoption and cybersecurity readiness. This gap is not a temporary delay but an institutional condition that arises when digitalisation advances faster than the laws, institutions and technical competencies required to secure it. This imbalance transforms cybersecurity from a discrete technical function into a systemic feature of digital governance. Building on strategic alignment and socio-technical systems theory, this paper introduces the *cyber gap assessment* (CGA), a conceptual framework designed to make this misalignment visible through analysis. CGA considers digital adoption and cybersecurity capacity as parallel yet interdependent trajectories, distinguishing between two complementary dimensions: *level gaps*, which measure the magnitude of divergence at a given time, and *pace gaps*, which capture the difference in their rates of change. Together, these parameters form a two-dimensional diagnostic that indicates whether states are converging or diverging in their capacity to align technological growth with institutional protection. A typology based on this structure identifies four configurations: high adoption/low capacity; low adoption/high capacity; high adoption/high capacity; and low adoption/low capacity. Each configuration reflects a unique sequence of digital reform, institutional design and governance logic. This analysis demonstrates that high level gaps correspond to structural exposure, where digital systems extend beyond protective reach, while sustained pace gaps generate compounding vulnerabilities across cyber-physical sectors such as energy, transport and e-government. Conversely, convergence across both dimensions signals institutional agility and anticipatory governance. By formalising these relationships, CGA reframes cybersecurity capacity as a co-evolving dimension of digital transformation rather than an ex-post control mechanism. This paper concludes that narrowing the cyber gap requires synchronised governance, adaptive regulation, secure-by-design infrastructures and investment in human capital. This study offers policymakers and researchers a reproducible framework to diagnose structural imbalances and align digital ambitions with resilience.

Keywords: Digital transformation, Cybersecurity, National cybersecurity, Cybersecurity capacity, Cyber gap, Digital adoption, Cybersecurity governance

1. Introduction

Digital transformation has become a defining feature of national development strategies. Governments are digitising services, firms are restructuring around data, and societies are becoming more dependent on interconnected digital infrastructures (Mettler et al, 2024). These changes have resulted in significant improvement in efficiency, accessibility and innovation (Xu et al, 2025). However, the rapid expansion of digital systems has not been matched by improvements in cybersecurity capacity, which have created a systemic imbalance that makes vulnerabilities structural rather than incidental (Forscey et al, 2022).

Many studies demonstrate that while governments and firms rapidly digitise, cybersecurity maturity remains formative, with institutional and social constraints inhibiting progress (Creese et al, 2021). Empirical research shows that although digital transformation boosts efficiency and innovation, it also expands the attack surface faster than technical preparedness can adapt (Catal et al, 2023; Saeed et al, 2023; Fan, 2024; Mirzaei, 2024; Balarabe, 2025; Brezavšček & Baggia, 2025).

This paper addresses this persistent misalignment by introducing the concept of the cyber gap, which captures the structural divergence between the pace of digital adoption and the evolution of cybersecurity capacity. Unlike measures of cyber readiness or maturity, the cyber gap emphasises the relational misfit between digital adoption and cybersecurity capacity. A state may advance rapidly in digitalisation yet remain insecure if protective institutions and practices fail to evolve at the same pace. Digital agendas often prioritise adoption while sidelining cybersecurity (Hurel 2022; Stewart, 2023), undermining resilience and trust at the organisational level and generating systemic risks at the national level that adversaries can exploit (Nye, 2022). To address these challenges, this paper pursues two objectives. First, it develops a conceptual framework and typology for understanding the cyber gap as a structural misalignment between digital adoption and cybersecurity capacity. Second, it proposes an operational diagnostic tool – the *cyber gap assessment* (CGA) – that translates this framework into measurable dimensions.

2. Conceptual Foundations

Studies on cybersecurity have long examined the imbalance between digital adoption and cybersecurity capacity through related concepts such as cyber readiness, maturity and resilience (Büyüközkan & Güler, 2025). Cyber readiness refers to the ability of states or organisations to prevent, detect and respond to cyber threats, often evaluated using indices or staged models (Von Solms & Van Niekerk, 2013; Lee et al, 2025). Cyber maturity frameworks conceptualise progression through predefined stages of institutional development, assuming a linear trajectory of improvement in protective capabilities (Brezavšček & Baggia, 2025). Cyber resilience frameworks often emphasise adaptive capacity following incidents, primarily focusing on post-incident response and recovery, with limited attention to proactive alignment or co-evolution (Dupont et al, 2023; Araujo et al, 2024; Teichmann & Sergi, 2025). They do not sufficiently explain how digital adoption can outpace cybersecurity capacity, leaving systemic vulnerabilities unaddressed. Table 1 summarises the primary cybersecurity constructs discussed in this section, highlighting how the cyber gap surpasses existing frameworks of readiness, maturity and resilience.

Table 1: Comparative overview of cybersecurity constructs

Construct	Focus	Approach	Limitation
Cyber readiness	Ability to prevent, detect and respond to cyber threats	Assessed as the absolute level of preparedness through audits or composite indices	Ignores temporal dynamics and pace of digital adoption (Von Solms & Van Niekerk, 2013; International Telecommunication Union, 2021)
Cyber maturity	Staged development of institutions, policies and controls	Evaluated through level-based or staged capability models	Assumes linear, static progression of capacity building (Brezavšček & Baggia, 2025)
Cyber resilience	Capacity to absorb, adapt and recover from cyber incidents	Measured by recovery and continuity outcomes following disruption	Focuses on post-incident adaptation with limited proactive alignment (Dupont et al, 2023; Araujo et al, 2024; Teichmann & Sergi, 2025)
Cyber gap (this study)	Structural misalignment between digital adoption and cybersecurity capacity	Conceptualised as relative divergence between digital adoption and cybersecurity capacity trajectories	

The term *cyber gap* denotes a structural divergence between how fast nations digitise and how effectively they build and sustain cybersecurity capacity. The cyber gap arises from persistent patterns in digital governance, where the pursuit of digital adoption consistently outpaces the development of cybersecurity measures. In this context, the cyber gap is defined as a structural misalignment between two interdependent national trajectories: digital adoption and cybersecurity capacity. Digital adoption refers to the expansion of infrastructures and services such as broadband, e-government, digital payments and digital platforms, which are common benchmarks in digital development (International Telecommunication Union [ITU], 2021; Organisation for Economic Co-operation and Development [OECD], 2023). Cybersecurity capacity refers to the development of laws, institutions, incident response mechanisms, technical standards and a skilled cybersecurity workforce, as reflected in frameworks such as the ITU Global Cybersecurity Index and national capacity-building models (ITU, 2021; Collett et al, 2021; Lee et al, 2025). Instead of focusing on inequality in access or competence, the cyber gap highlights a relational imbalance that can also leave digitally advanced countries vulnerable to systemic threats.

Empirical evidence supports this observation. Mettler et al (2024) show that national digital-transformation policies evolve unevenly across infrastructural and governance sectors, while cybersecurity frameworks frequently lag behind the expansion of digital services, particularly in emerging economies. The ITU Global Cybersecurity Index (2021) demonstrates that high levels of digital adoption do not always match high cybersecurity capacity, revealing a persistent decoupling between technological diffusion and institutional preparedness.

As Craigen et al (2014) argued earlier, the dominance of a technical view of cybersecurity has constrained its development, limiting interdisciplinary integration across technological, organisational and policy areas. This helps explain why cybersecurity capacity often evolves reactively within broader digital-transformation efforts. In this context, the cyber gap redefines cybersecurity as a dynamic misalignment rather than a static deficit of capacity.

While the concept of the cyber gap explains why divergence emerges, its distribution across countries is highly uneven. Some states advance rapidly in digital adoption but lag behind in security, whereas others prioritise cybersecurity at the expense of innovation and digital progress. These contrasting trajectories underscore the need for a systematic way to classify how adoption and capacity interact in practice.

3. Mapping the Cyber Gap: A Conceptual Typology

Although the typology is conceptual, similar patterns can be observed in global indicators such as the ITU Global Cybersecurity Index (2021) and the United Nations E-Government Development Index (2022), which together reveal how states differ in digital progress and their ability to secure it. By mapping these configurations, the typology explains how different national trajectories of digitalisation and cybersecurity capacity produce distinct vulnerability profiles. Such typological reasoning has long been used to identify structural patterns that quantitative metrics alone might miss (Doty & Glick, 1994; Elman, 2005). This section proposes a typology with four categories, offering a conceptual framework for comparative analysis.

3.1 High Digital Adoption/low Cybersecurity Capacity

The first quadrant represents the archetypal expression of the cyber gap. States in this quadrant undergo rapid digitalisation through expanded digital transformation efforts, such as e-government platforms, fintech ecosystems and data-driven public services, while their cybersecurity institutions remain comparatively underdeveloped. Political and economic incentives to modernise often drive technological deployment faster than governance reforms can keep pace with, resulting in a systemic misfit between two interdependent trajectories: technological expansion and capacity building (Pérez-Nordtvedt et al, 2008).

The technical layer of transformation, such as networks, platforms and applications, can scale rapidly, whereas the corresponding layers of policy design, enforcement and professional expertise evolve more slowly (Ansari & Garud, 2009). This generates governance lag as digital infrastructures expand while the rules, safeguards and human capabilities required to protect them remain incomplete (Őri & Szabó, 2024). Cybersecurity capacity does not erode through neglect but through asymmetric modernisation, a condition where innovation progresses faster than governance can stabilise (Thierer, 2018).

3.2 Low Digital Adoption/High Cybersecurity Capacity

At the opposite end of the spectrum are states that prioritise cybersecurity over digital expansion. They establish legal frameworks and specialised agencies early in their development, creating a governance structure focused on caution and control (ITU, 2021). This sequencing typically emerges in states with strong administrative traditions and centralised governance, where they value policy consistency more than rapid market experimentation. While this approach improves state preparedness, it also embeds caution – a core principle of digitalisation – that influences how innovation is subsequently authorised and scaled.

Institutional and protective capacity outpace innovation, creating rigidity rather than fragility as controls harden faster than use cases mature (Henderson & Venkatraman, 1993). Before the deployment of new technologies, rigorous compliance, risk assessment and security certification procedures often slow down adoption cycles. While these mechanisms improve reliability and reduce exposure, they hinder iterative experimentation. Past studies explain that formalising processes and controls enhances reliability but restricts exploratory activities, thereby reducing innovative change (Benner & Tushman, 2003; Cinar et al, 2022). Gradually, such environments cultivate risk aversion, causing organisations and public agencies to prioritise compliance over innovation. This imbalance delays the spread of digital services, constrains cross-sector collaboration, reduces flexibility and hampers adaptive learning. Consequently, exposure to external threats is lessened, but experimentation slows and digital transformation falls behind (Nelson & Madnick, 2017; Heierhoff & Hoffmann, 2022). However, recent empirical research shows a similar pattern: increased cybersecurity risks and precautions are associated with lower innovation intensity, underscoring the stability–agility trade-off at the core of this dynamic (Wang, Ho & Shan, 2024).

3.3 High Digital Adoption/High Cybersecurity Capacity

The third quadrant represents the closest alignment between digital adoption and cybersecurity capacity. States in this category integrate cybersecurity capacity as a core element of digital transformation rather than as a subsequent correction. Policy coordination, education systems and regulatory oversight develop alongside technological infrastructure, producing a dynamic equilibrium: the continuous co-adaptation of technical, institutional and human subsystems (Geels, 2002; OECD, 2023). In this quadrant, governance structures

proactively address risks instead of reacting to them, integrating cybersecurity principles into every stage of digital service design. Consequently, national strategies treat resilience as a capacity that evolves with technological ecosystems.

This category illustrates coherence across multiple levels of governance. Strategic intent, operational processes and technical capabilities are synchronised, reducing the gap between innovation and cybersecurity capacity (Henderson & Venkatraman, 1993; Chan & Reich, 2007). As digital services expand, cybersecurity frameworks mature in tandem, ensuring that both trajectories remain balanced over time (ENISA, 2020; National Institute of Standards and Technology [NIST], 2024).

However, this balance is dynamic rather than static. As digitalisation accelerates, maintaining alignment requires continuous adaptation and institutional agility, a capacity reflected in both national policy frameworks and international governance models (Teece, 2007; OECD, 2023; NIST, 2024). Therefore, the real marker of maturity in this quadrant is not the initial achievement of balance, but the ability to sustain and recalibrate it amid technological and organisational change. High adoption and high capacity do not eliminate the cyber gap but transform it into an ongoing challenge of governance, vigilance and renewal.

3.4 Low Digital Adoption/low Cybersecurity Capacity

This category describes states that remain at early stages of both digitalisation and cybersecurity development. Limited fiscal resources, institutional fragility and dependence on external technology providers constrain progress across both domains (Hurel, 2022). The cyber gap here is latent until digitalisation begins to accelerate.

In the short term, these countries appear less exposed because their digital footprint is small. However, this limited exposure can hide deep structural vulnerabilities, as underlying weaknesses in institutional design, education systems and digital infrastructure remain concealed beneath low connectivity rates. As digital adoption expands, often through e-government projects or imported platforms for finance, education or public administration, their cybersecurity capacity rarely scales at the same rate. Legal frameworks, enforcement mechanisms and skilled human resources tend to lag, leaving new systems dependent on external expertise and inconsistent regulatory protections (Collett, 2021; Hurel, 2022).

States with a dual constraint of low technological adoption and minimal cybersecurity readiness possess limited institutional capacity, which may postpone the emergence of the cyber gap but ultimately exposes a deferred vulnerability. Preventing this requires anticipatory governance that invests early in legal and institutional foundations, enabling cybersecurity capacity to evolve alongside future digital growth (OECD, 2023).

These four quadrants map the structural diversity of national manifestations of the cyber gap, capturing how states differ in sequencing, governance priorities and institutional adaptability. The typology demonstrates that imbalance is a relational condition arising from sequencing reforms, the design of institutions and the logic of governance that either links or decouples adoption from cybersecurity capacity. By translating the abstract notion of the cyber gap into these distinct national trajectories, the typology offers a comparative lens through which policymakers and scholars can diagnose how well the two trajectories co-evolve. The following section expands on this typology by exploring methods to measure and compare the cyber gap among states through the use of composite indicators and dynamic alignment metrics.

4. Conceptual Operationalisation of CGA

To make contrasts analytically usable, this paper introduces CGA as a conceptual diagnostic. Treating the two constructs of digital adoption and cybersecurity capacity as parallel but interdependent trajectories allows for a relational expression of the cyber gap rather than viewing it as an absolute measure. In practice, increased adoption may stimulate protective investments, while improved capacity can accelerate further adoption. However, for diagnostic clarity, this assessment treats them as analytically separate but empirically interdependent.

CGA distinguishes between two complementary dimensions:

- **Level gap (GL):** This represents the relative position of digital adoption and cybersecurity capacity at a given time.
- **Pace gap (GD):** This represents the difference in the year-on-year growth rates of digital adoption and cybersecurity capacity.

These two parameters form a two-dimensional vector (GL, GD) capturing both state and momentum. A positive GL indicates that adoption is ahead of capacity, while a positive GD shows that adoption is accelerating faster than capacity, as previously explained. This vector approach avoids reducing multidimensional complexity into a single composite score (Saisana, Saltelli & Tarantola, 2005) and instead offers a diagnostic view of relative trajectories.

Conceptually, CGA functions as a lens rather than a scorecard. Its diagnostic value lies in revealing where imbalance occurs and how it evolves while remaining modular across different indicator sets and country contexts (Creese et al, 2021). In policy terms, it bridges the typology and the pathways that follow: CGA bands map directly onto regulatory levers, secure-by-design policies, workforce development, operational readiness and international cooperation. By emphasising the parallel development of adoption and capacity, the CGA reinforces the principle of anticipatory governance, encouraging states to scale protective capacity before vulnerabilities mature into systemic risk (Collett et al, 2021).

Table 2: Conceptual components of CGA

Component	Description
Digital adoption (A)	Trajectory of digital infrastructures and services (e.g. broadband, e-government, digital payments)
Cybersecurity capacity (P)	Trajectory of cybersecurity capacity (e.g. laws, computer emergency response teams, standards, workforce, governance)
Level gap (GL)	Relative position of digital adoption versus cybersecurity capacity at a given time
Pace gap (GD)	Difference in year-on-year growth of digital adoption and cybersecurity capacity
Vector form	CGA = (GL, GD) captures both state and momentum.
Decision bands	Surge risk, security drag and convergence (based on persistence over multiple years)
Purpose	Diagnostic tool to link typology to institutional drivers and policy levers

5. Mathematical Representation of CGA

This formalisation follows established practices in composite-indicator design and enables a reproducible way to quantify the relative trajectories of digital adoption and cybersecurity capacity across states.

Let:

- $\mathrm{A}_{i,t}^{A}$ = digital adoption score for country *i* at time *t,*
- $\mathrm{P}_{i,t}^{A}$ = cybersecurity capacity score for country *i* at time *t.*

Because these indicators are heterogeneous in scale and distribution, they are standardised using a per-year z-score transformation (OECD & JRC, 2008):

$$\mathrm{z}_{i,t}^{A} = \frac{\mathrm{A}_{i,t} - \mu_t^{A}}{\sigma_t^{A}}, \mathrm{z}_{i,t}^{P} = \frac{\mathrm{P}_{i,t} - \mu_t^{P}}{\sigma_t^{P}},$$

where μ_t and σ_t represent the cross-sectional mean and standard deviation across all countries in year t. The level gap captures the static divergence between digital adoption and cybersecurity capacity:

$$\mathrm{G}_{i,t}^{L} = z_{i,t}^{A} - z_{i,t}^{P}.$$

A positive $G_{i,t}^{L}$ indicates that digital adoption exceeds cybersecurity capacity (structural exposure), while a negative value reflects a security-first orientation where cybersecurity capacity leads digital adoption.

The pace gap captures the dynamic divergence in year-on-year change:

$$\mathrm{G}_{i,t}^{D} = \left(\mathrm{z}_{i,t}^{A} - \mathrm{z}_{i,t-1}^{A}\right) - \left(\mathrm{z}_{i,t}^{P} - \mathrm{z}_{i,t-1}^{P}\right),$$

where a positive $G_{i,t}^{D}$ signals that digital adoption is accelerating faster than cybersecurity capacity. Both components together yield a relational representation of the cyber gap over time. To minimise volatility from single-year fluctuations, $G_{i,t}^{D}$ can be smoothed using a multi-year moving average (Chatfield, 2000). Moreover, country-level indicators may be subject to definitional inconsistencies or data gaps. Smoothing techniques and robustness checks can help mitigate such issues and ensure interpretive stability (Box et al, 2015).

These two components of the gap define the assessment vector:

$$\mathrm{G}_{i,t} = \left(\mathrm{G}_{i,t}^{L}, \mathrm{G}_{i,t}^{D}\right),$$

representing both the static alignment (level) and dynamic momentum (pace) of each country's digital trajectory.

However, the vector form remains the primary analytical lens, as it preserves state and direction, preventing the loss of interpretive richness that comes with aggregation.

- $G^L > 0 \rightarrow$ digital adoption exceeds cybersecurity capacity (*structural exposure*)
- $G^D > 0 \rightarrow$ digital adoption accelerates faster than cybersecurity capacity (*dynamic risk*)
- $G^L \approx 0$ and $G^D \approx 0 \rightarrow$ relative alignment (*convergence*)

By explicitly outlining this relational structure, CGA offers a transparent and replicable foundation for identifying areas of divergence and convergence between digital adoption and cybersecurity capacity, as well as determining where strategic intervention is most needed.

6. Strategic Implications of CGA

CGA demonstrates that misalignment between digital adoption and cybersecurity capacity is not merely conceptual but has systemic consequences. High level gap values reveal digitalisation advancing beyond protective capacity, translating directly into strategic vulnerability. Research on hybrid conflict shows that adversaries exploit weak protective capacity rather than technological inferiority (Hoffman, 2007; Rid, 2012). States with such imbalances are more susceptible to espionage, disruption and disinformation. As Cavelty and Egloff (2019) note, even advanced economies with sophisticated infrastructure remain strategically fragile when governance and enforcement lag.

When the pace gap persists, digital trajectories accelerate faster than protective capacity can adapt, compounding risk across security, infrastructure, trust and international relations. This pattern is particularly acute in cyber-physical domains such as energy grids, transport networks and e-health systems. Moreover, resilience theory warns that vulnerabilities in one subsystem can cascade across others (Perrow, 1999; Linkov et al, 2018). Evidence from ransomware attacks on critical infrastructure confirms how delayed cybersecurity capacity multiplies disruption (Benmalek, 2024). When acceleration continues unchecked, insecure systems become embedded, creating compounding risks and costly retrofits. Therefore, CGA links sustained pace gap trajectories with structural fragility in critical sectors that cannot easily be secured after deployment.

Public trust is central to digital governance. Empirical research shows that data breaches and misuse of information erode confidence, undermining adoption (Ou et al, 2022; Stewart, 2023). Within CGA, high level gap and persistent pace gap values indicate environments where protective institutions lag behind growth. Citizens face insecure e-government services, unprotected digital identities and poorly regulated fintech platforms. Thus, persistent misalignment generates a paradox: the drive for modernisation risks delegitimising the transformation when security fails to converge.

Finally, CGA reveals how divergence and convergence shape states' relative positions in the international system. Convergent states, where level gap and pace gap remain near zero, gain reputational advantages as trusted hubs for digital trade and governance (Madise & Martens, 2006). CGA highlights how structural misalignment crystallises globally, redistributing power along lines of security capacity.

This imbalance transforms cybersecurity from a technical function into a structural condition of digital governance. Bridging the cyber gap requires expanding capacity and sustaining synchronisation to ensure that institutional development keeps up with technological advancements. Hence, CGA is an analytical tool and a diagnostic lens that indicates how nations manage the balance between digital ambition and resilience.

7. Conclusion

This paper introduced the cyber gap as a structural misfit between the trajectories of digital adoption and cybersecurity capacity. Through CGA, the paper reconceptualised national readiness not as a fixed score but as a relational dynamic between two interdependent systems: digital adoption and cybersecurity capacity. The distinction between level gaps and pace gaps reveals how institutional inertia, workforce shortages, governance fragmentation and uneven international cooperation perpetuate imbalance over time. Practically, the paper provides policymakers with a diagnostic lens to locate vulnerabilities, anticipate systemic risks and sequence reforms that align digital expansion with protective capacity. While this study develops the cyber gap assessment

as a conceptual and analytical framework, its full potential will emerge through empirical application. Testing the CGA with multi-year datasets that combine indicators of digital adoption and cybersecurity capacity would enable validation of the typology and calibration of both level and pace gaps across countries. Future research can extend this work by operationalising the framework in comparative analyses to explore regional patterns, temporal dynamics, and the policy effectiveness of interventions aimed at narrowing the cyber gap. Bridging the cyber gap is a matter of governance design, ensuring that each advance in connectivity is matched by an equivalent advance in cybersecurity capacity.

Ethics Declaration: Ethical clearance was not required for this research, as it did not involve human participants, personal data, or experimental procedures. This research is based solely on conceptual analysis and secondary data from publicly available sources.

AI Declaration: No generative artificial intelligence (AI) tools were used in the conception, design, analysis, or writing of this paper. All content was created by the author. Standard software applications were used only for text editing and reference management.

References

Ansari, S. and Garud, R. (2009) 'Inter-generational transitions in socio-technical systems: The case of mobile communications', *Research Policy*, Vol. 38, No. 3, pp 382–392. https://doi.org/10.1016/j.respol.2008.11.009

Araujo, M.S.D., Machado, B.A.S. and Passos, F.U. (2024) 'Resilience in the context of cyber security: A review of the fundamental concepts and relevance', *Applied Sciences*, Vol. 14, No. 5, 2116. https://doi.org/10.3390/app14052116

Balarabe, A. (2025) 'Digital borders and beyond: Establishing normative grounds for cybersecurity and sovereignty in international law', *Computer Law & Security Review*, Vol. 53, 106321. https://doi.org/10.1016/j.clsr.2025.106321

Benmalek, M. (2024) 'Ransomware on cyber-physical systems: Taxonomies, case studies, security gaps, and open challenges', *Internet of Things and Cyber-Physical Systems*, Vol. 4, pp 186–202. https://doi.org/10.1016/j.iotcps.2023.12.001

Benner, M.J. and Tushman, M.L. (2003) 'Exploitation, exploration, and process management: The productivity dilemma revisited', *Academy of Management Review*, Vol. 28, No. 2, pp 238–256. https://doi.org/10.5465/amr.2003.9416096

Box, G.E.P., Jenkins, G.M., Reinsel, G.C. and Ljung, G.M. (2015) *Time Series Analysis: Forecasting and Control*. 5th edn. Hoboken, Wiley, New Jersey.

Brezavšček, A. and Baggia, A. (2025) 'Recent trends in information and cyber security maturity assessment: A systematic literature review', *Systems*, Vol. 13, No. 1, p 52. https://doi.org/10.3390/systems13010052

Büyüközkan, G. and Güler, M. (2025) 'Cybersecurity maturity model: Systematic literature review and a proposed model', *Technological Forecasting and Social Change*, Vol. 213, 123996. https://doi.org/10.1016/j.techfore.2025.123996

Catal, C., Ozcan, A., Donmez, E. and Kasif, A. (2023) 'Analysis of cyber security knowledge gaps based on cyber security body of knowledge', *Education and Information Technologies*, Vol. 28 No. 2, pp 1809–1831. https://doi.org/10.1007/s10639-022-11261-8

Cavelty, M.D. and Egloff, F.J. (2019) 'The politics of cybersecurity: Beyond the technical', *Global Policy*, Vol. 10, No. 3, pp 456–471.

Chan, Y.E. and Reich, B.H. (2007) 'IT alignment: What have we learned?', *Journal of Information Technology*, Vol. 22, No. 4, pp 297–315. https://doi.org/10.1057/palgrave.jit.2000109

Chatfield, C. (2000) *Time-Series Forecasting*. Boca Raton, Chapman & Hall/CRC, Florida.

Cinar, E., Demircioglu, M.A. and Yildiz, M. (2022) 'Public sector innovation in context: A comparative study of innovation types', *Public Management Review*, Vol. 26, No. 1, pp 265–292. https://doi.org/10.1080/14719037.2022.2080860

Collett, R. (2021) 'Understanding cybersecurity capacity building and its relationship to norms and confidence building measures', *Journal of Cyber Policy*, Vol. 6, No. 3, pp 298–317. https://doi.org/10.1080/23738871.2021.1948582

Craigen, D., Diakun-Thibault, N. and Purse, R. (2014) 'Defining cybersecurity', *Technology Innovation Management Review*, Vol. 4, No. 10, pp 13–21. Available at: https://timreview.ca/article/835

Creese, S., Goldsmith, M., Nurse, J.R.C. and Phillips, E. (2021) 'A cybersecurity capacity maturity model: Cross-national insights and lessons learned', *Journal of Cyber Policy*, Vol. 6, No. 3, pp 394–416. https://doi.org/10.1080/23738871.2021.1995811

Doty, D.H. and Glick, W.H. (1994) 'Typologies as a unique form of theory building: Toward improved understanding and modeling', *Academy of Management Review*, Vol. 19, No. 2, pp 230–251.

Dupont, B., Shearing, C., Bernier, M. and Leukfeldt, R. (2023) 'The tensions of cyber-resilience: From sensemaking to practice', *Computers & Security*, Vol. 132, 103372. https://doi.org/10.1016/j.cose.2023.103372

Elman, C. (2005) 'Explanatory typologies in qualitative studies of international politics', *International Organization*, Vol. 59, No. 2, pp 293–326. https://doi.org/10.1017/S0020818305050101

ENISA (European Union Agency for Cybersecurity) (2020) *National Cybersecurity Strategies: Practical Guide on Development and Execution*. Athens: ENISA. Available at: https://www.enisa.europa.eu/publications/national-cyber-security-strategies-toolset

Fan, P. (2024) 'Governance of high-risk technologies: The role of fragmented global regimes', *Global Political Economy*, Vol. 3, No. 2, pp 265–283. https://doi.org/10.1332/263514224X17211200207760

Forscey, D., Bateman, J., Beecroft, N. and Woods, B. (2022) *Systemic Cyber Risk: A Primer*, Carnegie Endowment for International Peace, Washington, DC.

Geels, F.W. (2002) 'Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case study', *Research Policy*, Vol. 31, No.8–9, pp 1257–1274. https://doi.org/10.1016/S0048-7333(02)00062-8

Heierhoff, S. and Hoffmann, N. (2022) 'Cyber security vs. digital innovation: A trade-off for logistics companies?', in *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS)*, pp 6792–6801. Available at: https://hdl.handle.net/10125/80161

Henderson, J.C. and Venkatraman, N. (1993) 'Strategic alignment: Leveraging information technology for transforming organizations', *IBM Systems Journal*, Vol. 32, No. 1, pp 4–16. https://doi.org/10.1147/sj.321.0004

Hoffman, F.G. (2007) *Conflict in the 21st Century: The Rise of Hybrid Wars*, Potomac Institute for Policy Studies, Arlington, Virginia.

Hurel, L.M. (2022) 'Interrogating the cybersecurity development agenda: A critical reflection', *The International Spectator*, Vol. 57, No. 3, pp 66–84. https://doi.org/10.1080/03932729.2022.2095824

International Telecommunication Union (ITU) (2021) *Global Cybersecurity Index (GCI) 2020*. Geneva: ITU. Available at: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2021-PDF-E.pdf

Lee, G., Kim, S., Lee, I., Brown, S. and Carbajal, Y.A. (2025) 'Adapting cybersecurity maturity models for resource-constrained settings: A case study of Peru', *The Electronic Journal of Information Systems in Developing Countries*, Vol. 91, No. 1, e12350. https://doi.org/10.1002/isd2.12350

Linkov, I., Trump, B.D., Poinsatte-Jones, K. and Florin, M.-V. (2018) 'Governance strategies for a sustainable digital world', *Sustainability*, Vol. 10, No. 2, p 440.

Madise, Ü. and Martens, T. (2006) 'E-voting in Estonia 2005: The first practice of legally binding Internet voting in the world', *Electronic Voting in Europe*, pp 15–26.

Mettler, T., Miscione, G., Jacobs, C.D. and Guenduez, A.A. (2024) 'Same same but different: How policies frame societal-level digital transformation', *Government Information Quarterly*, Vol. 41, No. 2, 101932. https://doi.org/10.1016/j.giq.2024.101932

Mirzaei, A. (2024) 'Cybersecurity governance: Fragmentation, coordination, and institutional challenges', *Journal of Cybersecurity*, Vol. 10, No. 1, taae012. https://doi.org/10.1093/cybsec/taae012

National Institute of Standards and Technology (NIST) (2024) *The NIST Cybersecurity Framework (CSF) 2.0*. Gaithersburg, MD: NIST. https://doi.org/10.6028/NIST.CSWP.29

Nelson, N. and Madnick, S. (2017) 'Studying the tension between digital innovation and cybersecurity', in *Proceedings of the 3rd International Conference on Information Systems Security and Privacy (SIGSEC)*, Porto, Portugal, 19–21 February.

Nye, J.S. Jr. (2022) 'Cyber power and the U.S. national security strategy', in Cox, M. and Stokes, D. (eds.) *The Routledge Handbook of U.S. Foreign Policy*. 2nd edn. Abingdon: Routledge, pp. 603–616

Organisation for Economic Co-operation and Development (OECD) (2023) *OECD Policy Framework on Digital Security: Towards Economic and Social Prosperity*, OECD Publishing, Paris. https://doi.org/10.1787/9789264276268-en

OECD and Joint Research Centre (JRC) (2008) *Handbook on Constructing Composite Indicators: Methodology and User Guide,* OECD Publishing, Paris. https://doi.org/10.1787/9789264043466-en

Őri, D. and Szabó, Z. (2024) 'A systematic literature review on business–IT misalignment research', *Information Systems and e-Business Management*, Vol. 22, pp 139–169. https://doi.org/10.1007/s10257-023-00664-w

Ou, C.X., Zhang, X., Angelopoulos, S., Davison, R.M. and Janse, N. (2022) 'Security breaches and organization response strategy: Exploring consumers' threat and coping appraisals', *International Journal of Information Management*, Vol. 65, 102498. https://doi.org/10.1016/j.ijinfomgt.2022.102498

Pérez-Nordtvedt, L., Payne, G.T., Short, J.C. and Kedia, B.L. (2008) 'An entrainment-based model of temporal organizational fit, misfit, and performance', *Organization Science*, Vol. 19, No. 5, pp 785–801.

Perrow, C. (1999) *Normal Accidents: Living with High-Risk Technologies*. 2nd edn. Princeton University Press, Princeton, New Jersey.

Rid, T. (2012) 'Cyber war will not take place', *Journal of Strategic Studies*, Vol. 35, No. 1, pp 5–32.

Saeed, S., Altamimi, S.A., Alkayyal, N.A., Alshehri, E. and Alabbad, D.A. (2023) 'Digital transformation and cybersecurity challenges for business resilience: Issues and recommendations', *Sensors*, Vol. 23, No. 15, 6666. https://doi.org/10.3390/s23156666

Saisana, M., Saltelli, A. and Tarantola, S. (2005) 'Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, Vol. 168, No. 2, pp 307–323. https://doi.org/10.1111/j.1467-985X.2005.00350.x

Stewart, H. (2023) 'Digital transformation security challenges', *Journal of Computer Information Systems*, Vol. 63, No. 4, pp 919–936. https://doi.org/10.1080/08874417.2022.2115953

Teece, D.J. (2007) 'Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance', *Strategic Management Journal*, Vol. 28, No. 13, pp 1319–1350. https://doi.org/10.1002/smj.640

Teichmann, F. and Sergi, B.S. (2025) 'The EU Cyber Resilience Act: Hybrid governance, compliance, and cybersecurity regulation in the digital ecosystem', *Computer Law & Security Review*, 59, 106209. https://doi.org/10.1016/j.clsr.2025.106209

Thierer, A. (2018) *The Pacing Problem and the Future of Technology Regulation,* Mercatus Center at George Mason University, Arlington, Virginia.

Von Solms, R. and Van Niekerk, J. (2013) 'From information security to cyber security', *Computers & Security*, Vol. 38, pp 97–102. https://doi.org/10.1016/j.cose.2013.04.004

Wang, J., Ho, C.Y. and Shan, Y.G. (2024) 'Does cybersecurity risk stifle corporate innovation activities?', *International Review of Financial Analysis*, Vol. 91, 103028. https://doi.org/10.1016/j.irfa.2023.103028

Xu, Y., Ji, J., Qiao, Y. and Huang, J. (2025) 'How and when does digital transformation promote technological innovation performance? A study of Chinese high-tech firms', *Technovation*, Vol. 146, 103294. https://doi.org/10.1016/j.technovation.2025.103294

Evaluating an Investigative Process for Cryptocurrency-Related Crimes

Johnny Botha[1], Kreaan Singh[1], and Louise Leenen[2]
[1]Council for Scientific and Industrial Research, Pretoria, South Africa
[2]University of Western Cape and CAIR, Cape Town, South Africa

jbotha1@csir.co.za
ksingh1@csir.co.za
lleenen@uwc.ac.za

Abstract: This paper evaluates a previously proposed investigative process for cryptocurrency-related crimes, originally introduced by the authors (Botha, Singh, & Leenen, 2025a), through the application of a real-world case study. The process covers crime reporting and case registration, on-chain analysis, off-chain analysis, and the transformation of investigative intelligence into court-admissible evidence. The current study focuses on a new, active case involving an elderly South African (SA) woman who was defrauded of a substantial portion of her pension through a fraudulent investment scheme known as ###-Platform (redacted). The case is presently under investigation by the Directorate for Priority Crime Investigation (DPCI), a specialised unit of the South African Police Services (SAPS) tasked with addressing serious economic crimes and commonly referred to as the Hawks. By systematically applying the proposed investigative process to this case, the study assesses the framework's practical utility, adaptability, and effectiveness in real-world conditions. The analysis further reflects on legal, technical, and procedural challenges encountered during the investigation, offering critical insights for law enforcement, regulators, and cybersecurity professionals. It also highlights broader systemic vulnerabilities that facilitate such scams, particularly among elderly and non-technical populations. The findings underscore the need for enhanced public education, improved regulatory oversight, and international cooperation in combating cryptocurrency fraud. Ultimately, the paper contributes to the evolving discourse on financial crime in the digital age and aims to support the development of more secure and accountable crypto-investment environments.

Keywords: Blockchain, Chain analysis, Cryptocurrency crime, Investment crypto-scam, OSINT

1. Introduction

The adoption of blockchain technology has surged in recent years, and with it, a rise in criminal activity whereby fraudsters and organised crime use the complexities of blockchains to mask their activities. This paper analyses a fraudulent cryptocurrency investment and trading scam by using a investigate process proposed in previous publications (Botha, Singh, & Leenen, 2025a). Cryptocurrency trading or investment scams typically begin on social media or through messaging apps. It should be noted that unsolicited contact from an unknown individual or an online acquaintance introducing an unfamiliar trading platform significantly increases the likelihood of fraudulent activity (CTFC, 2025).

This paper examines a case centered around a fraudulent investment platform called ###-Platform. The platform's website is currently inactive. The paper includes background details provided by the victim related to this case. DPCI, also known as the Hawks, in SA, is leading the investigation. The case adheres to the phases and steps outlined in the proposed methodology detailed in the journal article: 'A Proposed Bitcoin Blockchain Investigation Methodology: Based on a Case Study Approach' by (Botha, Singh, & Leenen, 2025a). Initially, the study presents the case study background, followed by an analysis of the case on and off the blockchain, according to the proposed process steps. Once sufficient intelligence is gathered for profiling the target entity, a link analysis is conducted. The final step involves law enforcement procedures. The study aims to assess or evaluate the process through a case study approach. Feedback on the process is provided, discussing its strengths, weaknesses, and potential improvements. The final section concludes the paper's findings.

2. Case Background

An elderly woman contacted one of the authors, J Botha, to assist in the investigation and analysis of a fraudulent case in which she fell victim to. The victim was persuaded to invest in ###-Platform, which was portrayed to be a global online trading entity based in the United Kingdom (UK), integrating traditional investments with crypto-based funding and trading mechanisms. The platform promised to manage elderly victims' pensions and to generate large profits from the proceeds. The company advertised leveraged trading across various assets, including Bitcoin (BTC), Ethereum (ETH), the S&P 500, and Tesla, amongst others. ###-Platform asserted that their company had experienced significant growth, served international clients, and accessed top-tier liquidity

and a diverse array of trading tools while ensuring security to facilitate a safe and efficient trading environment for all participants (Trustpilot, 2023).

In March 2023, a criminal case was opened and assigned to a law enforcement officer from DPCI to conduct the investigation. The first author, J Botha, interacted with the official investigator to gather all necessary information for assistance. The victim had opened an account with VALR, a cryptocurrency exchange based in SA. It is important to note that VALR has no connection to ###-Platform. The victim went through customer onboarding, providing all necessary Know Your Customer (KYC) documentation and bank statements as required by VALR. She was granted an account, but while depositing money to VALR, VALR identified unusual deposit patterns considering the customer's risk profile and age. Upon inquiry, the customer assured VALR that she was not receiving assistance from anyone to perform these deposits, was not being scammed and thus requested proper allocation of the deposits. This attestation was accepted by VALR and the victims deposit of SA Rands (ZAR) were unrestricted. She acquired cryptocurrency and transferred her funds to ###-Platform. Once the victim realised that she was being defrauded, her responses to the prior questions changed significantly. She informed VALR of her use of ###-Platform and the substantial returns that were promised. Additionally, the victim revealed that the scammers had accessed her personal computer through the AnyDesk software application. AnyDesk is a remote desktop application that allows someone to take control over if permission is granted (AnyDesk, 2025). According to the victim, the scammers purportedly executed trades on her behalf on ###-Platform, as she was unfamiliar with the procedure.

The victim supplied the cryptocurrency address used in VALR from where the transactions were executed. She claimed to have lost close to ZAR4,950,000 (US $281,312) over two years. The claims correspond to the bank statements provided by the victim, showing deposits made into VALR from her bank account. It also corresponds with the police statement.

3. Analysis of the Case

The investigative process previously proposed by (Botha, Singh, & Leenen, 2025a) serves as the foundation for the analysis presented in this paper. The process has five phases, namely:

- Data collection
- Analysis
- Theory development and validation
- Suspect identification and reasonable grounds
- Legal action

The investigative process is depicted in Figure 1. For the full details of the process and steps, refer to the journal paper (Botha, Singh, & Leenen, 2025a). The analysis of the case study in this paper, follows this process and its proposed steps, which is outlined in the next sections with comprehensive explanations of each step with the case information provided as input.

3.1 Opening of Case

The opening of a case forms part of the data collection phase. In this particular use case, an investigation is already underway. Therefore, the first step to open a case has already been done. The victim has provided a BTC cryptocurrency address that was linked to the crypto exchange VALR in SA. Furthermore, a compilation of names and phone numbers of suspect individuals who contacted the victim and eventually defrauded her are also available. The victim provided the following inputs *(redacted)*:

- VALR transaction History File with a number of addresses and transactions. The two main in this file addresses were:
- *32jC##…QaYG*
- *1Lac##…XFpU*
- Web domain
- *www.###-platform.io*
- Names
- *List of names of people communicating over WhatsApp and Phone calls (multiple people and name – probably all fake names)*
- Phone Numbers
- *List of numbers that were calling the victim (SA and UK numbers)*

- Email Addresses
- *s#####e@cryptodotcom.info*
- *List of scammer names@###-platform.io*
- Physical Address
- *# B### St, London E## #BG, UK*

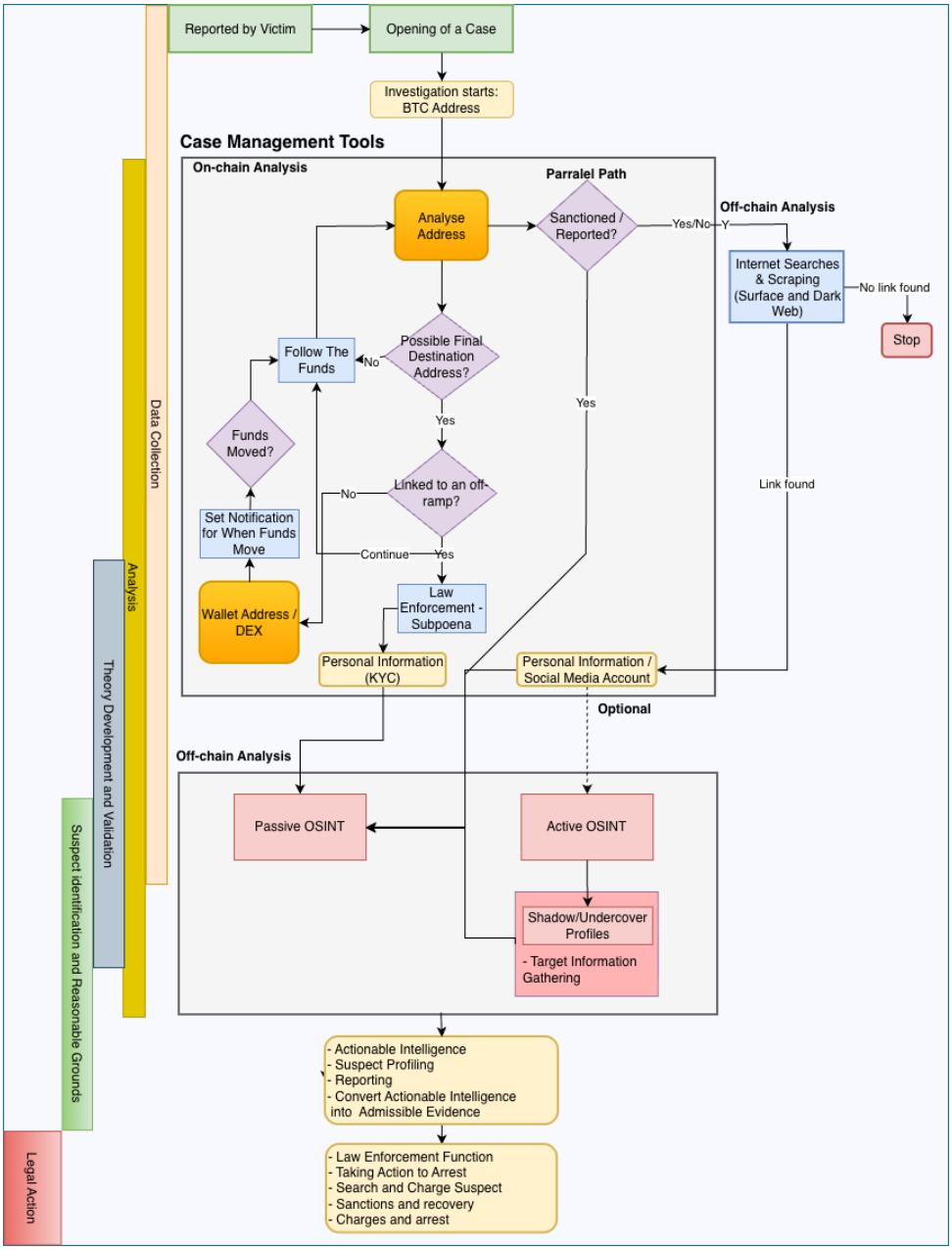


Figure 1: Process for Conducting Cryptocurrency Crime Investigations (Botha, Singh, & Leenen, 2025a)

3.2 On-chain Analysis

The on-chain analysis is primarily conducted during the analysis phase, but it also intersects with the data collection as well as the theory development and validation phases (refer to Figure 1). As the investigation advances, additional data is gathered, and new theories are continually developed and validated. The blockchain analysis tool known as Breadcrumbs (Breadcrumbs, 2023) was used to perform the on-chain analysis. This type of analysis examines blockchain data, focusing on transaction history, wallet activities, and smart contract interactions between users (TransFi, 2025) (Yang, Klages-Mundt, & Gudgeon, 2023). The analysis of this particular case began with the victim's two cryptocurrency addresses found in the transaction history file. The file contained crypto transactions, and detailed payments made to VALR that were converted into BTC from the victim's account. The address 32jC##…QaYG found in the file is the victim's BTC receiving address. There was

also one outgoing transaction from the victim's BTC address 32jC##…QaYG to a crypto BTC wallet address 1Lac##…FXpU. Because the victim's address was associated with VALR, the tool couldn't track outgoing transactions from the exchange. Blockchain analysis tools cannot follow the funds through a centralised exchange (CEX) because the exchange has internal transactions that occurs off-chain within its private internal ledger, not on the public blockchain. This breaks the public traceability (Chainalysis.com, 2020). However, with law enforcement's help, a section 205 subpoena was issued to VALR, leading them to disclose the outgoing transactions to the scammer's addresses. A section 205 subpoena is a legal tool under the SA Criminal Procedure Act that allows authorities to compel individuals or entities to provide evidence of information relevant to a criminal investigation. This includes ordering a third party, like a cryptocurrency exchange in this case, to disclose information and records (The Banking Association Aouth Africa, 2024). It was discovered that the scammers transferred the stolen cryptocurrency to three different addresses (see Table 1) including the reference names for each scammer address used in this paper.

Table 1: Scammer Addresses

Scammer Address	Reference Name in Paper
1End##…4EFo (which still holds 1.14 BTC)	Scammer-Address-1
142T##…xmGo (0.28 BTC)	Scammer-Address-2
1NvB##…gHfD (12.8 BTC)	Scammer-Address-3

VALR also provided a list of outgoing transactions from the scammer's addresses that linked to multiple cryptocurrency exchanges outside of SA. These transactions, addresses and linked exchanges are illustrated in Table 2. The flow of these transactions is shown in Figure 3.

Table 2: Outgoing Transactions from Scammer Addresses

TX	Address	Exchange
Bctq##…9zwn	32Eq##…ubrF	Coinpayments.net
bc1q##…qd0t	33jD##…Jen2	Kyrrex.com
bc1q##…xvw5	3Q7h##…izCB	Kyrrex.com
bc1q##…4eea	3AXC##…3Gvh	Kyrrex.com
bc1q##…jsnw	31ik##…8a6J	Kyrrex.com
bc1g##…4vm3	3HG7##…xRRu	Kyrrex.com
bc1g##…dmgy	33jD##…Jen2	Kyrrex.com

Another discovery was an additional VALR address, 39F4##…jVDX, linked to Scammer Address 3. Intelligence gathered from VALR led to the disclosure of the name and surname behind the address. This address was found to be the actual address belonging to the victim. The name behind the address found in the transaction history file, was the name of her daughter's former boyfriend who, according to the victim, was the one who introduced her to the investment platform. The victim stated that the former boyfriend had also fallen victim to the fraudulent platform, however he has not opened a criminal case to investigate the matter. More research is needed to accurately determine the ex-boyfriend's role and link to the victim and scammers addresses. In this paper, the victim's actual address will be referred to as 'Victim-Address-1' and the address of the former boyfriend as 'Victim-Address-2'.

Figure 2 illustrates the transaction flow originating from Victim-Address-2, the address of the former boyfriend of the victim's daughter and the victim's address, Victim-Address-1, into to the three scammer addresses. Three transaction flows will be discussed here, for each one of the scammer cryptocurrency addresses from the list above, to several cryptocurrency exchanges, based on the information obtained from the section 205 subpoena on VALR. The three scammer addresses are shown on the right side of Figure 2, inside the rectangle.

3.2.1 Scammer-Address-1

No direct transaction was identified with Scammer-Address-1. However, the funds were transferred from VALR, Victim Address 1, to a deposit address, 1LvG##…7Xpi, on Binance exchange, the largest cryptocurrency exchange globally. A section 205 subpoena was requested on this address to obtain the account holder details. However, no results have been obtained yet, at the time of writing. From Binance, the funds moved to an external wallet, bv1q2h…edux. Afterward, the funds travelled through seven transactions and cryptocurrency wallet addresses

to reach Scammer-Address-1, which currently still holds 1.14 BTC. This address is marked as a destination address, since no further transactions took place. A monitoring notification was placed using the tool Breadcrumbs, to notify the analyst via email or SMS when the funds are on the move again. If the address is proven to be linked to criminal proceeds, the address should be reported to sanctioning bodies, but, as an interim measure, it has been reported on chainabuse.com.

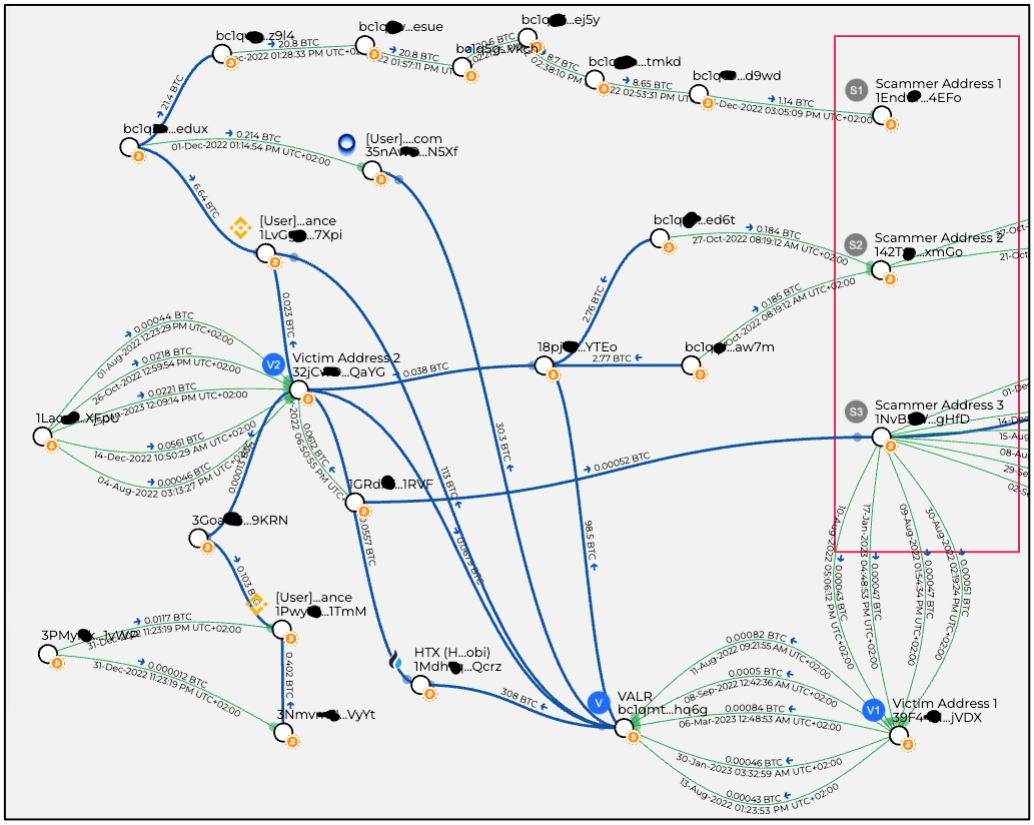


Figure 2: Outgoing Transactions from Victim Addresses to Scammer Addresses

3.2.2 Scammer-Address-2

Funds were transferred into Scammer-Address-2, see Figure 2 (middle right side of the figure) and Figure 3 (top left of the figure) through another wallet address as well as via the exchange HTX, also known as Huobi (HTX, 2025). From Scammer-Address-2, funds were transferred into two cryptocurrency wallet addresses, bc1q##…qqdd and bc1q##…fg8l, as illustrated in Figure 3 (top left and top middle side of the figure). Subsequently, these funds were moved to the exchange kyrrex.com, arriving at the address 3AXC##…3Gvh (top right side of Figure). The address in question is designated as a final destination due to the inability to track funds further. Nonetheless, according to trustworthy intelligence, 30,655.78 USDT was withdrawn from the exchange kyrrex.com into the external address TGXF##…MEVb, associated with the Tron blockchain. From this point, the funds were moved to other external Tron crypto addresses, including a transaction to a Binance address, see Figure 4. A section 205 subpoena is necessary to acquire details about the account holder and outgoing transactions from Binance. As for the additional crypto wallet addresses, further tracing of funds will be done to identify more destination addresses and exchanges.

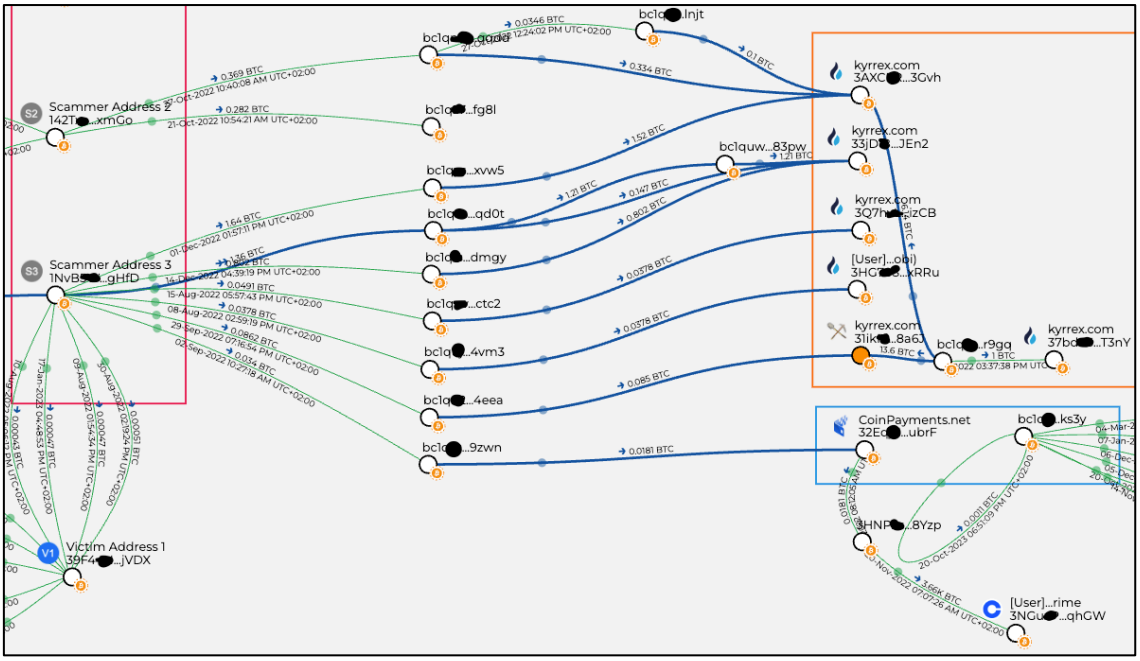


Figure 3: Outgoing Transactions from Scammer's Addresses to Cryptocurrency Exchanges

3.2.3 Scammer-Address-3

From the section 205 subpoena results received on VALR, a number of transactions flowed out of Scammer-Address-3 into the exchange kyrrex.com. These addresses can be seen in Figure 4 on the left side, inside the square. The Breadcrumbs tool flagged the address 31ik##…8a6J as a high-risk address (marked with a darker dot on the address node at the bottom left in Figure 4 and can also be seen in Figure 3, middle right inside the rectangle), meaning it has been linked to scams before. A section 205 subpoena is necessary to obtain further outgoing transactions and information on the account holder. Until this is issued, no further tracing can be performed. However, reliable information for Scammer-Address-3 (refer to Figure 3) also included a list of transactions out of these kyrrexx.com addresses to external addresses. Funds were converted from BTC into USDT on the Tron blockchain, and transferred to two Tron addresses, refer to Figure 4 (on the right).

Still following the path from Scammer-Address-3, based on the information from the VALR section 205 subpoena, transactions were made into the exchange CoinPayments.net (refer to Figure 3). The Breadcrumbs tool revealed that payments were made from CoinPayments.net to an external wallet and then into the exchange Luno. A subpoena is required to obtain further information beyond the Luno transfer. The same intelligence received for the previous address also revealed that from the payments made into CoinPayments.net, the change was paid into a BTC address also linked to CoinPayments.net. Note that the change refers to the Unspent Transaction Outputs (UTXOs). UTXOs are split into the payment to a receiver, and the change is then returned to the sender (ankura, 2024). The remaining BTC was then sent to external wallet addresses, of which some still hold BTC. One of these wallet addresses is the address bc1q##…ks3y (see Figure 5, top left inside the rectangle). From this wallet, traces of payments into the exchanges Coinbase, Binance, HitBTC and ByBit were found (refer to Figure 5). Through law enforcement, a section 205 subpoena is required to obtain account information and outgoing transactions at these exchanges.

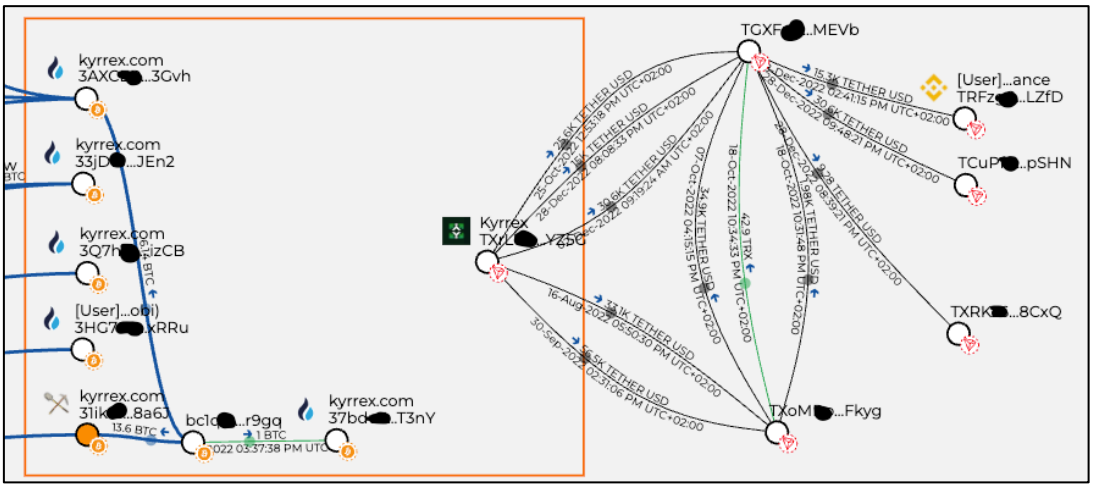


Figure 4: Outgoing Transactions From kyrrex.com

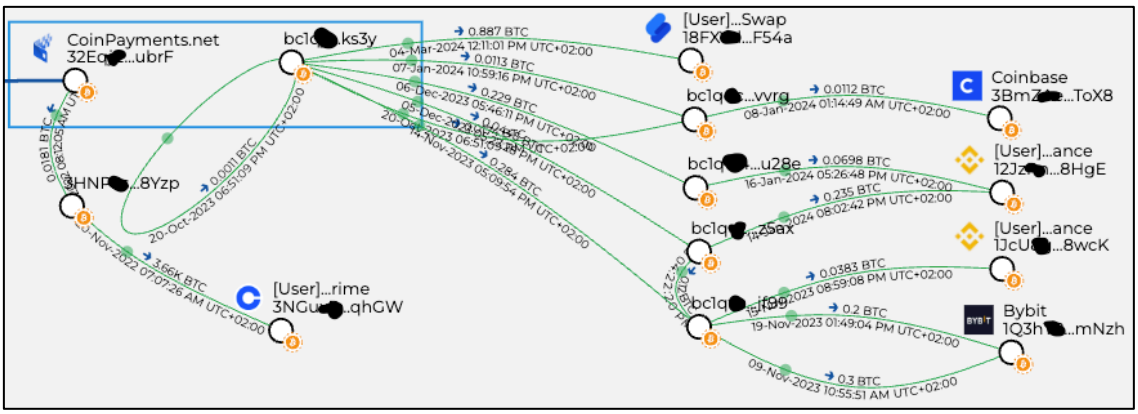


Figure 5: Outgoing Transactions from CoinPayments.net

3.3 Off-Chain Analysis

The off-chain analysis is also included in the analysis phase and overlaps with both the data collection phase and the theory development and validation phases. As the investigation moves forward, additional data and information are uncovered and gathered, while new theories are being developed and validated, refer to Figure 1.

3.3.1 Parallel path

The objective of this path is to search for a connection between each crypto address and an online presence of the address, whether it be on social media, web pages or the like.

Google searches were done on the three scammer addresses but returned no results. The AI Chat Bot for Google, Gemini (Google Gemini AI Chat Bot, 2025), was also used to see if any results could be retrieved. It returned a result that the addresses were flagged on chainabuse.com as involved in fraudulent activities. However, this was linked to the report that was filed by the analyst. Therefore, no positive results were found. This method is also categorised as Open-source Intelligence (OSINT), which is covered in the following section.

3.3.2 OSINT

OSINT involves gathering, processing, and correlating information that is publicly accessible from open data sources like social media, forums, blogs, public government data, publications, or commercial data. OSINT methodologies begin with initial input data and employ advanced collection and analysis techniques to progressively enhance understanding about a target individual or organisation. As each new piece of information is discovered, analysts move closer to their ultimate objective (Pastor-Galindo, Nespoli, Gómez Mármol, & Martínez Pérez, 2020).

In this segment of the analysis, OSINT methods are used to attempt to reveal the suspect's identity, based on the details provided by the victim as described in Section 3.1. During this phase, data collection continues as

new information is discovered online. The gathered data and information are transformed into actionable intelligence, leading to the suspect's profiling. Finally, the actionable intelligence should be converted into evidence that is admissible in court.

- Passive OSINT

Passive OSINT is the process of gathering information from publicly available sources such as Google, Whois records, Netcraft, books, articles, and similar open data repositories (Kolenbrander, 2025). In the case of an investigation, passive OSINT analysis aims to collect personal data about a target without any direct interaction with them. This method utilises various tools to find specific details, aimed at constructing a profile of the targetß individual (Botha, Singh, & Leenen, 2025a).

List of Names and Phone Numbers

The victim provided a list that included names and phone numbers from SA) and the UK. Through Truecaller (Truecaller, 2025), it was discovered that the SA numbers are linked to the ###-platform ZA Office and that these numbers are Voice-over-IP (VoIP). An additional check was conducted on claritycheck.com (a paid service for finding information on phone numbers and email addresses) (ClarityCheck, 2025), yielding no meaningful results. Searches on Google were carried out for names associated with ###-platform, and each name was connected to negative reviews on hellopeter.com and Trustpilot. The names are not allowed to be revealed in this paper. What can be disclosed is that the names are typical white male English speaking South African names, and can also be typical UK names.

Domain Name

The domain's expiration has resulted in a lack of available information online using ViewDNS.Info (ViewDNS.info, 2025). While the Wayback Machine (Wayback Machine, 2025) offered some insights, the domain's active period was brief. The domain lookup tool, who.is, provided no details, only confirming the domain's non-existence. None of the domain tools uncovered any useful data. According to the Gemini AI tool, the domain has minimal online presence, raising concerns due to the lack of basic information. Gemini suggested that the platform might be fraudulent with the intent to steal money, and it strongly recommended avoiding any interaction with it (Gemini, 2025).

On scam detector (scam-detector.com, 2025), a website that provides original information, reporting, reviews and analysis on websites, domains and e-commerce platforms, and advising if these are legitimately, safe and trustworthy, the ###-platform.io web domain got a score of 40.7% which signals risky and red flags. The domain was created on 23 June 2022, and the victim was scam started in December 2022. Shortly after the domain was created. Scam detector also indicated it has a proximity to suspicious websites of 73% and a phishing score of 78%. The website's meta data was poorly configured which would not help its online presence and loses credibility. The domain was active for 1 year, 11 months before it became inactive. Running the web domain through scam adviser (SCAMADVISER, 2025) it gave a very low trust score of 17%.

The domain www.###-platform.io was associated with negative feedback on hellopeter.com (hellopeter.com, 2025) and trustpilot.com (Trustpilot, 2023). Notably, the names and numbers given by the victim also appeared in complaints from other victims on these platforms, suggesting that the scammers targeted multiple individuals using the same identifiers. Some complaints noted that the scammers previously operated under different domain names, rs-fx.com and eduprocentre.com, indicating a prolonged period of fraudulent activity which continues with the current domain.

Another search result from Google pointed to the website www.55brokers.com, which suggests that the company ###-platform appears to operate as a UK-based entity, with the authority to provide services in the UK. Nonetheless, its regulation is solely in St. Vincent and the Grenadines, classifying it as offshore. Therefore, this is typically indicative of a scam (55BROKERS, 2025).

Email Address

The provided email address, s######s@cryptodotcom.info was subjected to various verification tools, revealing that the email account does not exist.

Regarding the names list email addresses all belonging to the domain ###-platform.io, various user complaints were found on websites like ScamAdviser, Trustpilot and Reddit, involving unsolicited emails promoting 'guaranteed' investment opportunities or recovery services for lost funds. It has been reported as 'pig butchering' or advance-fee scams, where scammers build trust via email or chat, then request deposits or

personal information. Pig butchering is an investment scam where fraudsters gain the trust of victims of time and then deceive them into investing in fake investment platforms (often fake cryptocurrency investment platforms) (DFPI, 2025).

No credible business or professional profile for any of the names were found to be linked to these email accounts in legitimate directories like LinkedIn or corporate registries.

- Active OSINT

Active OSINT typically involves the use of a program or script to collect data and leaves a log behind (Kolenbrander, 2025). Active OSINT analysis also entails interacting with a target person under deceptive circumstances, such as using a fake or undercover profile, to persuade them into disclosing personal information. This approach, however, carries a higher likelihood of the target discovering the subterfuge, potentially resulting in them changing their behaviour. It is important to emphasize that, without explicit authorization from a law enforcement body, conducting active OSINT is illegal in most jurisdictions. If such permission is granted, the information collected through active OSINT can later be used in a passive manner by scanning online platforms to gather fragmented information about the target (Botha, Singh, & Leenen, 2025a). For this investigation, no active OSINT methods were utilised. It could be considered in the near future with the legal permissions granted.

3.4 Link Analysis

Link analysis is the identification of relationships and connections between two entities such as individuals, organisations, etc (Holzer, Dietz, & Yang, 2016). This technique visually represents interconnected entities as nodes and links, facilitating a clearer understanding of the data and its underlying structure (Cambridge Intelligence, 2025). The link analysis enabled visual mapping of the connections between the victim and the suspect, including cryptocurrency transactions. By employing OSINT techniques on the data provided by the victim, a number amount of relevant information was successfully gathered. However, not enough data could be retrieved to successfully identify a target suspect. When collecting information on a target, a case management strategy is needed to keep track and make sense of the findings. For this case, the link analysis and case management tool used was Maltego (see Figure 6). Link analysis remains a component of the analysis stage, as well as of the theory formulation and verification stage. However, the main focus is placed on the suspect identification and reasonable grounds phase, refer to Figure 1.

Referring to Figure 6, the top left section illustrates the victim connected to her mobile number, email address, and a link to ABSA bank. Funds were moved from ABSA bank to the cryptocurrency exchange VALR. The BTC address associated with VALR is the victim's BTC deposit address at VALR. The three BTC addresses connected to the victim's address are those of the scammers, to which funds were transferred from the victim’s address and account at VALR. This data was acquired through a subpoena by law enforcement. The remainder of the blockchain link analysis is not included in this graph, as it was already illustrated in detail using the Breadcrumbs tool, as shown in figures 2, 3, 4 and 5. The three scammer addresses are connected to the entity ###-Platform. All information available on the platform is linked to this entity, including the domain name, email addresses, lists of phone numbers from the UK and SA, and a list of names (likely pseudonyms) of individuals who contacted the victim. WhatsApp chat messages between the victim and suspects have been retrieved and stored in Maltego. However, these chats did not yield further insights into the scammers’ personal details beyond the fake names used. Connections were made to the websites hellopeter.com and trustpilot.com. Numerous negative reviews were found and associated with the platform name, and the names from the list provided by the victim also appeared in these negative reviews.

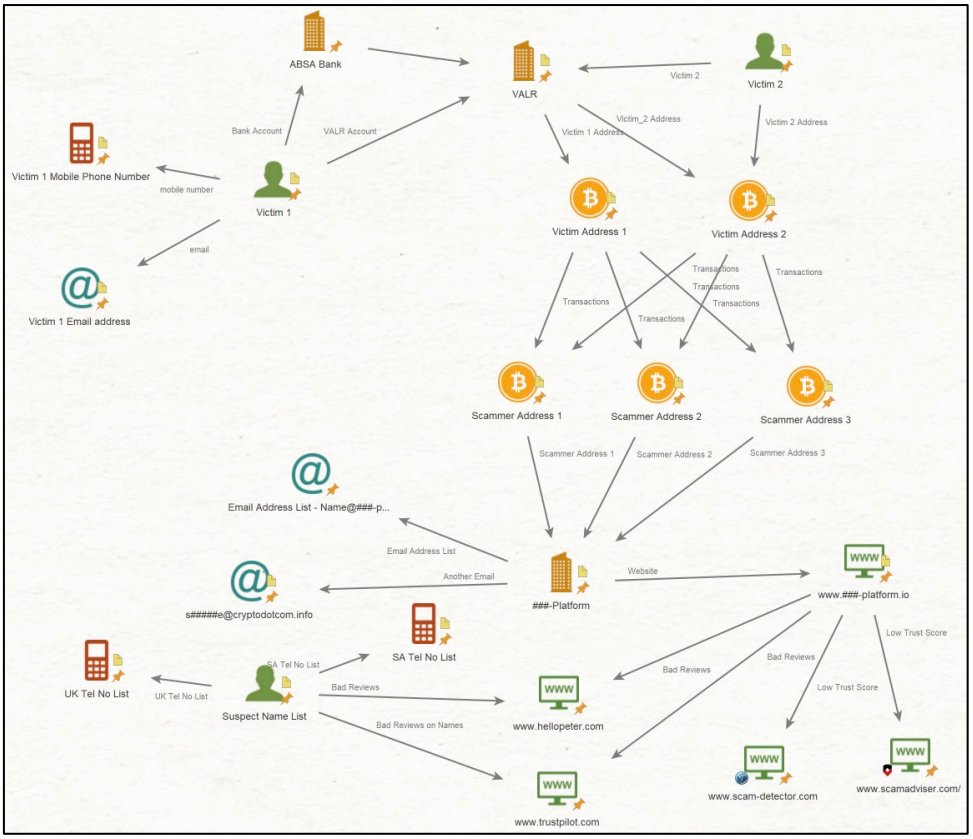


Figure 6: Link Analysis

3.5 Law Enforcement Function

This segment of the investigation is part of the legal action phase, refer to Figure 1, and is conducted by legal authorities to locate and prosecute the suspect, initiating the process of charging, arresting, imposing sanctions, recovering assets, and seizing funds.

In this study, certain legal actions were performed. The cryptocurrency exchange VALR received a subpoena and was instructed to disclose the personal information of the account holder along with all outgoing transactions. This information was obtained and proved to be highly beneficial for this investigation. Additionally, a subpoena was sent to the exchange kyrrex.com to provide the personal details of the account holder and outgoing transactions, but no information has been received yet. Nevertheless, through a law enforcement contact, intelligence was able to acquire the outgoing transactions. The account holder's user name was also obtained; however, it was not useful as the registered name was 'M Mpayme', which is a made-up name.

Unfortunately, the identities of the scammers remain undisclosed, and this phase cannot be fully executed without sufficient intelligence and evidence to proceed. Scammers are sufficiently adept at hiding their traces. The best hope is to follow the funds on the blockchain, conducting on-chain analysis, with the hope that the scammers make a mistake and leaks certain information that could be used in the off-chain analysis. A follow up study will be done in the near future that will focus mainly on the off-chain analysis. Ultimately the goal is to obtain the identity of the scammers and bring them to justice.

4. Process Evaluation

By employing a case study approach, the process can be tested and refined in practical scenarios. Previous studies conducted by the authors focused on a number of cases that has also follow the proposed process and with each case the process was progressively refined. The initial development of the process can be found in the analysis of the MTI Crypto Investment Scam" (Botha, Pederson, & Leenen, 2023). Another example involved an investigation into the fraudulent use of BTC in a divorce proceeding (Botha & Leenen, 2024). The same principles were applied in this case, particularly in relation to tracing funds on the blockchain. Following the second case, the process was further refined in another study, which introduced a high-level methodology describing how OSINT is applied in blockchain investigations (Gertenbach, Botha, & Leenen, 2024). A third incident involving a

cryptocurrency giveaway scam was investigated and followed the proposed process. This process was then refined once more in a subsequent study. (Botha, Singh, & Leenen, 2025a). Another case on a crypto fake investment scam was done following the refined process (Botha, Singh, & Leenen, 2025b). The specific case examined in this paper followed the refined process with the aim of determining where it could be further improved.

The investigation process offers a structured and systematic method for tackling cryptocurrency crime investigations. It underscores the necessity of collecting data from diverse sources to construct a comprehensive view of illegal activities. A significant emphasis is placed on clustering and pattern analysis, fundamental techniques in blockchain analytics, crucial for revealing the identities behind cryptocurrency transactions. The process also highlights the importance of link analysis and visualisation, which are critical for understanding complex transaction networks and for communicating findings to stakeholders.

The process is crafted to be practical and relevant for real-world investigations by law enforcement, financial institutions, and other organisations. OSINT techniques are crucial in this approach for associating blockchain activities with a particular individual or entity. The off-chain section of the process does not provide enough detail. Enhancements could include offering specific guidance and additional details on the techniques employed at each stage. While the on-chain analysis is detailed, the off-chain analysis and OSINT sections could benefit from more depth and examples. Examples of suspicious transaction patterns should be highlighted, and metrics on how the process was evaluated should be included. As blockchain technology and related investigative methods are continually advancing, the proposed process should be regularly updated to incorporate these developments.

5. Conclusion

With the growing adoption of blockchain technology, there's a corresponding increase in crimes and scams related to cryptocurrency. This paper offers practical insights by evaluating a structured investigative process that was previously proposed by the authors, relevant to law enforcement, regulators, investigators, analysts and researchers. The evaluation is done by taking a case study and follow the steps and phases proposed by the previously proposed methodology. This particular case study involves an elderly woman who was deceived by a fraudulent investment scheme. The case is currently under investigation by the CSIR (Council for Scientific and Industrial Research) in collaboration with South African law enforcement. No personal details of the victim have been disclosed in this paper. All cryptocurrency addresses and transactions have been masked to prevent complete identification of the addresses and transactions. The perpetrators persuaded the victim to transfer money using BTC, facilitated through the BTC blockchain. The proposed methodology or process is divided into five distinct phases, each accompanied by several steps. These steps often span multiple phases, meaning some steps might be part of two or three phases simultaneously. The steps include initiating a criminal case, conducting on-chain and off-chain analysis, and running a concurrent process for each cryptocurrency address to search for matches across social media platforms, forums, blogs, dark web markets, and more. OSINT techniques are employed, and link analysis is performed to visualise all collected data, making it understandable for external parties like law enforcement agencies. The information collected is transformed into actionable intelligence, which is then turned into evidence that is admissible in court. The final step involves legal procedures to prosecute and apprehend the suspect or target. The document offers a comprehensive assessment of the proposed procedure by analysing a specific case, highlighting its strengths and suggesting areas for improvement. Online scams continue to pose a major global threat due to the absence of standardised and official laws. Additionally, the borderless nature of cryptocurrency transactions has led to a rise in these scams within financial and cyber-crimes. In addition to evaluating the proposed process for investigating cryptocurrency crimes, this paper also aims to increase awareness about fake investment scams and fraudulent cryptocurrency investment platforms.

Ethics Declaration: Ethical clearance was not needed for this research.

AI Declaration: The academic AI tool Writefull was used to improve my English in the paper. The Gemini AI chat bot was used to see if I could find additional information on the scammer names, phone numbers and email addresses.

References

55BROKERS. (2025, Oct 20). *Is RSI-FX a scam or legit?* Retrieved from www.55brokers.com: https://55brokers.com/trader-inquiry/mr-51

ankura. (2024, Apr 25). *Crypto Asset Investigations: Key Considerations and Pitfalls*. Retrieved from ankura: https://angle.ankura.com/post/102j5sp/crypto-asset-investigations-key-considerations-and-pitfalls

AnyDesk. (2025, Oct 31). *Home Page*. Retrieved from anydesk.com: https://anydesk.com/en

Botha, J., Pederson, T., & Leenen, L. (2023). An Analysis of the MTI Crypto Investment Scam: User Case. *European Conference on Cyber Warfare and Security* (pp. 89-99). Athens, Greece: Academic Conferences International Limited.

Botha, J., & Leenen, L. (2024). Cryptocurrency-crime investigation: Fraudulent use of bitcoin in a divorce case. *International Conference on Cyber Warfare and Security (ICCWS)* (pp. 34-42). Johannesburg: Academic Conferences International Limited.

Botha, J., Singh, K., & Leenen, L. (2025b). Analysis of a Cryptocurrency Investment Scam: Pig Butchering. *European Conference on CyberWarfare and Security (ECCWS)* (pp. 61-70). Kaiserslautern: Academic Conferences International Limited.

Botha, J., Singh, K., & Leenen, L. (2025a, Feb 3). A Proposed Bitcoin Blockchain Investigation Methodology: Based on a Case Study Approach. *Journal of Information Warfare, 24*(1), 1-18. Retrieved Feb 10, 2025, from https://www.jinfowar.com/journal/volume-24-issue-1/proposed-bitcoin-blockchain-investigation-methodology-based-case-study-approach

Breadcrumbs. (2023, Oct 11). *Breadcrumbs Investigation*. Retrieved from https://www.breadcrumbs.app/: https://www.breadcrumbs.app/home

Cambridge Intelligence. (2025, Feb 10). *Link analysis*. Retrieved from www.cambridge-intelligence.com: https://cambridge-intelligence.com/why-link-analysis/

Chainalysis.com. (2020, Oct 9). *Why You Can't Trace Funds Through Services Using Blockchain Analysis (And Why You Don't Need to Anyway).* Retrieved from Chainalysis.com: https://www.chainalysis.com/blog/blockchain-analysis-trace-through-service-exchange

ClarityCheck. (2025, Oct 18). *Reverse Phone Lookup*. Retrieved from claritycheck.com: https://claritycheck.com

CTFC. (2025, Jan 08). *10 Signs of a Scam Crypto or Forex Trading Website*. Retrieved from ctfc.gov: https://www.cftc.gov/sites/default/files/2023-04/SpotFraudSites.pdf

DFPI. (2025, Oct 31). *Pig butchering – how to spot and report the scam*. Retrieved from Department of Financial Protection & Innovation (DFPI): https://dfpi.ca.gov/news/insights/pig-butchering-how-to-spot-and-report-the-scam/

Gemini. (2025, Oct 28). Gemini Chat Bot. *Information on rsi-platform.io*.

Gertenbach, W., Botha, J., & Leenen, L. (2024). A Proposed High-Level Methodology on How OSINT is applied in Blockchain Investigations. *International Conference onCyber Warfare and Security (ICCWS)* (pp. 75-83). Johannesburg: Academic Conferences International Limited.

Google Gemini AI Chat Bot. (2025, Oct 26). *https://chaton.ai/gemini/*. Retrieved from https://chaton.ai/gemini/: https://chaton.ai/gemini/

hellopeter.com. (2025, Oct 12). Retrieved from hellopeter.com: https://www.hellopeter.com/rsi-platform-daniel-miller

Holzer, C., Dietz, J., & Yang, B. (2016). Employing Link Analysis for the Improvement of Threat Intelligence Regarding Advanced Persistent Threats. *5th IAJC/ISAM International Conference* (p. 11). Orlando, Florida: International Association of Journals & Conferences (IAJC).

HTX. (2025, Oct 23). *Home Page*. Retrieved from HTX: https://www.htx.com/

Kolenbrander, J. (2025). Privacy Research using Active OSINT Techniques (p32). Virgina: Ph.D. thesis, Virgina Tech.

Pastor-Galindo, J., Nespoli, P., Gómez Mármol, F., & Martínez Pérez, G. (2020). The Not Yet Exploited Goldmine of OSINT: Opportunities, Open Challenges and Future Trends. *IEEE Access, 8*, 10282-10304.

SCAMADVISER. (2025, Oct 26). *SCAMADVISER*. Retrieved from https://www.scamadviser.com/check-website/rsi-platform.io

scam-detector.com. (2025, Oct 10). *Scam Detector Validator*. Retrieved from scam-detector.com: https://www.scam-detector.com/validator/rsi-platform-io-review

The Banking Association Aouth Africa. (2024, Dec 11). *Press Briefing Notes-Enhancing-DPCI-Capabilities.* Retrieved Oct 31, 2025, from banking.org.za: https://www.banking.org.za/wp-content/uploads/2024/12/press-briefing-notes-enhancing-dcpis-capabilities.pdf

TransFi. (2025, Feb 13). *What is On-Chain Analysis in Blockchain and How Do You Use It?* Retrieved from www.transfi.com: https://www.transfi.com/blog/on-chain-analysis-in-blockchain

Truecaller. (2025, Feb 10). *Number Search Results Page*. Retrieved from www.truecaller.com: https://www.truecaller.com/

Trustpilot. (2023, Oct 18). *RSI-Platform.io Review*. Retrieved from Trustpilot: https://www.trustpilot.com/review/rsi-platform.io

ViewDNS.info. (2025, Oct 7). *Home Page*. Retrieved from ViewDNS.info: https://viewdns.info/

Wayback Machine. (2025, Oct 25). Retrieved from Wayback Machine Internet Archive: https://web.archive.org/

Yang, Z., Klages-Mundt, A., & Gudgeon, L. (2023). Oracle Counterpoint: Relationships between On-chain and Off-chain Market Data. *arxiv*. Retrieved from https://arxiv.org/abs/2303.16331

Social Media's Role in Cyber Warfare: Attack Vectors, Space Ecosystem Integration, and Impacts

Gregory Broomfield
Marymount University, Arlington, USA

gbroomfi@marymount.edu

Abstract: Prompted by confusion surrounding several recent cyber incidents, this study examines how quickly social media has shifted from basic communication to a primary tool of modern cyber conflict. Platforms such as X (formerly Twitter), Telegram, TikTok, and others are rapidly being repurposed for sophisticated propaganda, targeting infrastructure, and concealing hostile motives behind waves of deliberately orchestrated false information. The Russia–Ukraine war, for example, demonstrates how fluid and adaptive these influence operations have become. Recent events have highlighted persistent vulnerabilities linked to growing connections between ground-based and satellite systems. Recent analyses indicate that the targeting of satellite-linked messaging services is more direct than previously recognized, and that AI-based botnets and deepfakes are proliferating rapidly in contemporary digital environments. This interdisciplinary study utilizes comparative case analysis and qualitative document analysis, integrating examples from the United States, European Union, China, Russia, and India to assess the evolving impact of hybrid space-cyber warfare. Theoretical models such as the Space-Cyber Hybrid Attack Matrix are referenced to categorize emergent threats, while global regulatory environments are contrasted to highlight transnational data governance challenges. Methodological rigor is supported by triangulating empirical data from documented incidents, governmental reports, and published cybersecurity expert testimony. Examination of international case studies and published accounts from cybersecurity professionals indicates that government responses are struggling to keep up, especially as disinformation spreads across borders at a daunting pace. Drawing from global case studies, this paper argues for the urgent need to rethink international cyber standards and adapt space-cyber treaties for robust security. The optimal path forward is through stronger coordination between the public and private sectors to effectively meet threats now manifesting at the intersection of social media algorithms, AI-generated influence operations, and expanding orbital communications ecosystems in both peace and conflict.

Keywords: Hybrid space-cyber warfare, AI-driven disinformation, Deepfakes, Satellite-linked messaging services, Transnational data flows, Hybrid warfare

1. Introduction

The rapid expansion of social media has reshaped contemporary cyber operations, turning digital networks into both battlefields and amplifiers of geopolitical conflict. Platforms such as X, Telegram, and TikTok now enable malicious actors to attack critical infrastructure, propagate false information, and exploit weaknesses in interconnected terrestrial and orbital networks. For example, platforms like X, Telegram, and TikTok are often used for psychological manipulation and influence operations (Mulahuwaish et al., 2025).

The 2022 Russia–Ukraine war showed how cyber, information, and space domains can converge. Military and non-state actors used social media, unencrypted satellite communications, and AI-driven botnets in coordinated campaigns (Iskoujina and Roberts, 2024; Li, 2023). As deepfakes and algorithmic disinformation spread, hurting both public trust and social cohesion (Pennycook et al., 2021), international policy frameworks have a hard time keeping up with new technologies and changing tactics. Moreover, these issues are exacerbated by international data flows and fragmented sovereignties, complicating both reaction and accountability.

Scholars such as Rahimi and Jones (2025) have argued that the networked nature of modern war means that cyber operations are not only tactical tools, but strategic levers in shaping everything from diplomatic signaling to public opinion. Cyber and space infrastructure are now so intertwined that attacks targeting one domain can cascade rapidly into the other, in processes known as "cross-domain synergy" (Clark, 2021). The resilience of societies against these threats increasingly depends not only on technical defenses, but on adaptable regulatory frameworks, multi-sectoral cooperation, and international coordination. The convergence of information technologies with satellite-linked systems and AI-driven content creation marks a critical juncture for global policy and national security, especially for attribution and deterrence (Brandqvist, 2024).

This paper aims to bridge critical research gaps by providing an interdisciplinary analysis of emergent attack vectors at the intersection of social media and space-based systems. This paper makes three principal contributions: (1) it introduces an expanded Space-Cyber Hybrid Attack Matrix to conceptualize cross-domain threats; (2) it synthesizes multi-region case studies on social media–enabled hybrid warfare; and (3) it advances a set of integrated policy recommendations for international cyber and space governance. The focus is both empirical and conceptual: examining case studies in the United States, European Union, China, Russia, and India;

theorizing hybrid risk through models such as the Space-Cyber Hybrid Attack Matrix; and developing explicit policy recommendations for a new era of global cyber governance. The central research questions are:

- How are adversaries weaponizing social media and satellite ecosystems to expand hybrid attack surfaces?
- What frameworks, detection mechanisms, and transnational policies best mitigate these converging threats?

2. Background and Literature Review

2.1 Hybrid Warfare: Expanding Domains

Hybrid warfare is characterized by deliberate synchronization among military, cyber, and informational spheres, integrating violent operations with psychological manipulation, economic extortion, and technical sabotage (Hoffman, 2007). Digital platforms not only facilitate real-time intelligence and influence but also hide attribution, so complicating accountability (Clark, 2021). Commercial satellite data and open-source intelligence have enhanced the velocity and accuracy of operations, enabling non-state players to engage in war at levels formerly restricted to nation-states.

China and Russia, for example, have made hybrid techniques part of their military doctrine. They use state-owned media and cyber proxies to compete strategically (Bendett, 2023). The connection between space and cyber is becoming more important, with attacks on satellites causing problems in energy grids, communication networks, and emergency services (Mulahuwaish et al., 2025). Being able to damage orbital infrastructure with either physical or cyber methods is now considered as a measure of a country's strength and resilience.

2.2 Social Media as an Attack Vector

Social media platforms have emerged as central tools for operational planning, influence, and execution in hybrid conflicts. Their capabilities for rapid amplification, audience segmentation, and viral dissemination make them ideal for coordinated attacks (Tucker et al., 2018; Vavaroutsos, 2025). Research by Tucker et al. (2018) demonstrates that targeted propaganda on social media can reliably change individual attitudes and shape population-level responses.

The rise of influencer networks , including those mobilized by adversarial states, enables the creation of "information bubbles" that can be seeded with false narratives, conspiracy theories, and divisive content (Marwick and Lewis, 2017). In the case of the Myanmar crisis, for example, Facebook was used systematically to incite ethnic violence and organize mass mobilization, with limited regulatory response (Mozur, 2018).

AI-driven bots, coordinated networks, and deepfakes have dramatically increased the scale of attackable vectors. Recent studies indicate that up to 70% of trending content in political hashtags during high-stakes events originates from botnets or coordinated networks rather than organic users (Ferrara et al., 2020).

The capacity for adversaries to manipulate the trending algorithms and mass-reporting features of platforms constitutes both direct and indirect threat vectors to democratic governance and crisis management.

2.3 Satellite Ecosystems and Space-Cyber Integration

Satellites and other orbital assets are very important for both military and civilian communications, navigation, remote sensing, and worldwide data transmission. These systems have weaknesses that go beyond just hacking or jamming. Increasingly, adversaries are using the connection between social media and land-based networks to create network-bound consequences. Peer-reviewed research indicates enduring vulnerabilities in satellite networks, such as unencrypted downlinks, authentication deficiencies, and the potential for adversaries to insert or intercept both telemetry and user data (Rahimi and Jones, 2025; Mulahuwaish et al., 2025).

The rise of small satellites and commercial space services has made the attack surface bigger, which means that supply chain risks and chances for enemies to take advantage of unpatched software, poor authentication methods, or old encryption are now greater (Foust, 2019; Rahimi and Jones, 2025). The connection between ground-based digital infrastructure and space-linked assets is now so strong that a breach in one area can swiftly spread to the other, turning local problems into worldwide ones. The lack of uniform technical standards and regulatory consistency makes it further harder to be resilient and respond (OECD, 2023).

2.4 AI-Driven Disinformation and Deepfakes

Disinformation efforts have shifted from manually disseminated false news to AI-enabled campaigns that use bots, language models, and synthetic media to mimic real sources, shape perceptions, and confuse detection systems (Pennycook et al., 2021). The psychological effect is made worse by advanced targeting algorithms that now divide people into groups based on their emotional triggers, belief systems, and social networks to get the most people to see and keep them.

Deepfakes are becoming increasingly realistic, which poses significant threats to the legitimacy of leaders, crisis communications, and trust in institutions. Synthetic video and audio can be used as weapons in high-profile negotiations, disasters, or political campaigns, as shown by cases in Taiwan, the US, and the EU (Brandqvist, 2024). Even though detection has gotten better, deepfakes are still hard to stop ahead of time. This means that more money needs to be spent on watermarking, digital provenance, and quick content flagging systems (Ferrara, 2020).

2.5 Transnational Data Flows and Regulatory Fragmentation

Global digital ecosystems depend on uninterrupted data flows, yet legal mandates for privacy, security, and localization generate fragmentation and technical obstacles. The GDPR sets high bars for data integrity and consumer protection, while China's Cybersecurity Law and Multi-Level Protection Scheme emphasize sovereign control and restricted sharing (DataGuidance, 2019; Cooley, 2022). The US continues to pursue sectoral approaches, resulting in gaps exploitable by adversaries.

Cloud services, multinational networks, and satellite relays routinely traverse legal jurisdictions, creating problems for joint incident response, evidence collection, and prosecution (Pillsbury, 2021; PwC, 2021). The World Economic Forum and OECD have called for harmonized frameworks, but progress is uneven (OECD, 2023). This regulatory divergence impedes the cross-border sharing of threat intelligence and technical best practices in cases of hybrid attacks.

2.6 Synthesis

Despite extensive scholarship on cyber and information attacks, there remains a gap in integrated policy analysis that specifically addresses the cross-domain complexity of social media and space-cyber threats. This study applies a holistic lens, connecting technical, organizational, and regulatory perspectives to advance both theory and actionable recommendations.

3. Methodology

This research adopts a qualitative, document-based methodology, integrating a systematic literature review with a comparative case study framework to analyze the hybridization of social media-driven cyber warfare and space-linked systems. Data were sourced from leading academic databases, including IEEE Xplore, ScienceDirect, and the ACM Digital Library, supplemented by major policy journals and institutional publications spanning 2017–2025. The inclusion criteria prioritized peer-reviewed, empirical, and international studies addressing social media manipulation, AI-driven disinformation, satellite-enabled cyber operations, and transnational regulatory responses.

To maximize empirical depth, the review also encompassed authoritative grey literature, such as government white papers, cybersecurity advisories, and officially documented incident reports from agencies including the World Economic Forum, OECD, and UN-affiliated organizations. Each case selected was evaluated on its relevance to the hybrid space-cyber paradigm, the breadth of documented empirical details, and the accessibility of technical and policy data across multiple jurisdictions.

Thematic synthesis and framework analysis were employed as core analytical strategies. Specifically, coded themes were derived from case and incident documentation and subsequently mapped to the Space–Cyber Hybrid Attack Matrix to assess emergent patterns, cross-domain vulnerabilities, and regulatory responses. Triangulation of findings was achieved through corroboration among independent studies, technical advisories, and multi-source policy documentation. Reliability was further supported by a rigorous selection process balancing sector, geography, and documented detail.

No primary data collection or interaction with human subjects was conducted. The research remained entirely within the bounds of secondary analysis of public and published materials.

4. Case Studies

The following case studies are interpreted using the Space-Cyber Hybrid Attack Matrix introduced in Section 7.1, with selected incidents explicitly mapped to vector–target–mechanism cells.

4.1 Russia–Ukraine War: Social Media and Hybrid Tactics

The Russia–Ukraine war represents one of the most thoroughly documented instances of hybrid warfare, where cyber, space, and information operations are deeply intertwined (CIGI, 2015; CyberPeace Institute, 2025). Russian actors launched waves of phishing and malware, coordinated across X, YouTube, and Telegram, while Ukrainian defenders relied on rapid open-source intelligence sharing and satellite-enabled civilian communication (Brandqvist, 2024). Disinformation campaigns focused on undermining morale, obscuring objectives, and sowing doubt about leadership capacity.

Throughout multiple invasion phases, deepfakes purporting to show Ukrainian officials capitulating or misdirected Russian units appeared within civilian messaging apps and news cycles, generating confusion and amplifying psychological stress (Pennycook et al., 2021). Ubiquitous use of bots and algorithmic content amplification reshaped the information environment, making rapid verification and counter-messaging a persistent challenge.

Technical analysis during the war revealed several instances in which satellite-linked messaging services were compromised, leading to the exposure of military movements, logistical planning, and sensitive communications. Defensive countermeasures such as switching satellite frequencies, employing multi-layer encryption, and mass reporting of suspicious content were partially effective but constrained by regulatory gaps and interoperability barriers.

For example, during the initial phases of the Russia–Ukraine war, compromises of satellite-linked messaging services exposed troop movements and logistics planning while disinformation campaigns simultaneously targeted civilian perceptions. This incident illustrates how adversaries can blend battlefield sensing, operational security degradation, and psychological pressure in a single operation.

This incident can be situated within the Space-Cyber Hybrid Attack Matrix as follows:

- Vector: Combined (space + cyber + information)
- Target: Command-and-control and civilian population
- Mechanisms: Satellite-link compromise (space), credential theft/malware (cyber), coordinated disinformation and panic-inducing narratives on social platforms (information)

4.2 Satellite Messaging and Data Leakage

Recent peer-reviewed literature and technical reports confirm that satellite network vulnerabilities, including persistent unencrypted downlinks, authentication flaws, and inadequate incident response, have enabled adversaries to intercept, manipulate, and disrupt both military and commercial systems (Rahimi and Jones, 2025; Mulahuwaish et al., 2025).

These breaches highlight systemwide deficiencies in prevailing encryption standards and the complexity of international coordination needed for remediation. The findings from (Rahimi and Jones, 2025) and (Mulahuwaish et al., 2025) demonstrate how researchers have accessed sensitive control signals and user data with relatively modest ground station equipment and make the urgency of cross-jurisdictional remediation and technical standardization clear.

Satellites have also offered unique opportunities for bypassing censorship or executing direct-to-populace propaganda campaigns, enabling the broadcast of strategic messaging to populations otherwise shielded by national controls. Such operations blur lines between technical sabotage and psychological warfare, requiring defensive responses both in orbit and on the ground (Foust, 2019).

4.3 AI-driven Disinformation Campaigns and Deepfakes

Synthetic media, AI-powered bots, and automated content amplification have become key pillars of modern hybrid operations. During the 2020–2023 election cycles in the US, France, and India, botnets generated massive volumes of deepfake audio and video, targeting political figures, journalists, and voters. These attacks, combined with panic-mongering and viral conspiracy theories, undermine democratic processes and increase polarization (Pennycook et al., 2021).

Analysis of platform logs revealed that deepfake content often originated from overseas servers, further complicating attribution and response. Published expert accounts describe the frustration of counter-disinformation efforts, noting that by the time fake content was flagged, retractions were far less visible than the original posts. This asymmetry between attack and defense was exacerbated by the lack of standardized AI moderation and limited cross-platform cooperation (Ferrara et al., 2020; Marwick and Lewis, 2017).

Regulatory interventions such as the EU's Digital Services Act and the US Honest Ads Act have begun addressing some gaps in platform accountability, but global harmonization remains a distant goal. Several private-sector initiatives, including the Global Disinformation Index and the Deepfake Detection Challenge, show promise but require further international investment and benchmarking.

The Myanmar crisis offers a contrasting case in which the primary operational vector was informational: coordinated Facebook campaigns amplified hate speech, dehumanizing narratives, and targeted calls to violence against Rohingya communities, with limited technical sophistication but devastating social impact.

This case maps within the Space-Cyber Hybrid Attack Matrix as:

- Vector: Information
- Target: Social trust and civilian population
- Mechanisms: Algorithmic amplification of inflammatory content, coordinated hate campaigns, offline mobilization triggered by online narratives.

4.4 Transnational Data Flows and Regulatory Barriers

The digital and orbital integration of infrastructure means that cyber incidents can traverse legal borders, undermining local control and complicating response. Major cross-border breaches over the last decade often crossed satellite links, cloud servers, and messaging platforms operating in multiple regulatory environments. The Facebook-Cambridge Analytica scandal and China's strict data localization requirements are prominent examples highlighting the difficulties in balancing openness, security, and sovereignty.

Published accounts from multinational cyber response teams revealed persistent obstacles in intelligence sharing, evidence preservation, and coordinated forensics. This was pronounced in space-cyber incidents, given the absence of clear jurisdiction assignment for assets or actions that span national borders. They emphasized the need for standardized attribution protocols and joint training exercises to bridge legal and procedural gaps (OECD, 2023).

This case maps within the Space-Cyber Hybrid Attack Matrix as:

- Vector: Cyber (legal/regulatory constraints on data flows)
- Target: Incident response and evidence chains (institutional resilience)
- Mechanisms: Data localization mandates, conflicting breach-notification rules, cross-border transfer restrictions affecting forensics

From a matrix perspective, divergent data-protection regimes manifest as cyber-vector constraints on incident response and forensic capability, indirectly expanding adversarial freedom of maneuver in cross-border operations.

4.5 Additional Vulnerable Sectors

Healthcare, education, and the energy sector have become frequent targets for social engineering and hybrid attacks leveraging social media amplification. Ransomware incidents in the UK's National Health Service and US academic networks reveal layered vulnerabilities stemming from poor security hygiene, inadequate training, and rapid digital adoption during pandemics (Rahimi and Jones, 2025; Brandqvist, 2024).

Simulation-based awareness programs, particularly in healthcare and education, have shown measurable benefits for operational resilience, reducing incident rates where implemented (Marsh-Armstrong et al., 2024). Nevertheless, these gains are uneven across regions, with resource-poor institutions still struggling against the volume and sophistication of emergent attack vectors.

5. Discussion

5.1 Complexity and Convergence

Examination of documented case studies and published literature reveals that adversary entities leverage multi-domain tactics, integrating cyber infiltration, psychological manipulation, and orbital system disruption (Clark, 2021; Mulahuwaish et al., 2025). Attribution is further hampered by intentional proxy utilization, falsified credentials, and the worldwide scope of digital networks (Bendett, 2023). Defensive capabilities are constrained by disjointed incident response teams, variable international standards, and outdated systems poorly suited to swiftly changing hybrid threats (OECD, 2023).

Technological innovations such as cyber-physical honeypots and AI-enhanced incident response demonstrate potential for anticipatory defense, as evidenced by institutional warnings and peer-reviewed research (Mulahuwaish et al., 2025; Rahimi and Jones, 2025). However, implementation encounters substantial obstacles in coherence, investment, and legal clarity, indicative of regulatory misalignment noted in international cybersecurity reports (OECD, 2023).

5.2 Psychological and Social Impacts

The literature provides substantial evidence of societal repercussions, particularly with the erosion of public trust, civic engagement, and institutional legitimacy in communities affected by hybrid and misinformation efforts (Marwick and Lewis, 2017; Pennycook et al., 2021). Notable studies indicate how misinformation and deepfake-induced terror dramatically alter public health behaviors, electoral turnout, and crisis management, ultimately reducing policy resilience and national identity (Rahimi and Jones, 2025).

Simulation and modeling studies demonstrate that population-level resilience to hybrid campaigns is significantly associated with digital literacy and trust in authorities (Pennycook et al., 2021; Brandqvist, 2024). Documented participatory counter-disinformation initiatives have resulted in enhancements; however, top-down messaging alone has resulted in few advantages (Rahimi and Jones, 2025).

5.3 Institutional Response and Policy Gaps

Policy analysis of government agencies, consortia, and private organizations highlights persistent difficulties in harmonizing international regulation and incident response, particularly regarding satellite encryption, social media authentication, and AI disinformation detection (Clark, 2021; OECD, 2023). Fragmented regulatory approaches and inconsistent threat intelligence sharing remain critical barriers to systemic resilience.

Resource constraints, policy inertia, and rapid technology cycles mean that most defense postures are largely reactive. Although multinational cyber defense initiatives and informational security working groups (e.g., NATO, UN) have made incremental gains, their effectiveness is limited by voluntary participation and diverging national interests, as assessed in global cybersecurity outlook reviews (OECD, 2023).

6. Policy Implications and Recommendations

A coherent response to hybrid space–cyber threats and AI-enabled information operations requires a forward-leaning policy agenda that is both technically robust and institutionally coordinated. This section sets out five interlocking recommendations that move from hardening orbital and digital infrastructures, to constraining AI-driven manipulation, to building the legal, governance, and organizational foundations needed for sustained resilience. Together, updating space–cyber treaties and technical standards, countering AI-driven disinformation at scale, fostering trusted global data governance, strengthening public-private collaboration, and enhancing cross-sectoral awareness and resilience provide a mutually reinforcing framework for managing the security, stability, and accountability of the emerging hybrid battlespace.

1. Update Space-Cyber Treaties and Technical Standards: Existing treaties must reflect the realities of cross-domain hybrid warfare, incorporating security standards for satellite-linked communications, shared monitoring, and incident reporting. Mechanisms for rapid technology standardization and joint enforcement must be prioritized by multilateral organizations such as the ITU and UN. An emphasis on end-to-end encryption, zero-trust architecture, and international audit capabilities is necessary for orbital resilience.
2. Counter AI-Driven Disinformation at Scale: Mandate social media and messaging platforms to develop and deploy real-time AI-fact-checking, deepfake detection, and cross-platform content verification. Enhance interoperability of detection frameworks and create sectoral benchmarks for platform

accountability. For maximum impact, agencies should incentivize open sharing of best practices and establish periodic regulatory review, leveraging international expertise (Pennycook et al., 2021).

3. Foster Trusted, Harmonized Global Data Governance: Advance multilateral data governance regimes, drawing on the successes of the GDPR and OECD evidence-based frameworks for global interoperability (OECD, 2023). Create new incentive structures for legal recognition and technical standardization, focusing on cross-border threat intelligence, digital forensic sharing, and satellite forensics. Develop and pilot international legal templates for hybrid attack attribution and evidence preservation.
4. Strengthen Public–Private Sector Collaboration: Build sustainable, multi-stakeholder channels for threat intelligence exchange, modeling on successful sectoral response networks, cyber ranges, and cross-sector simulations. Encourage investment in interdisciplinary training, public awareness, and rapid response protocols that integrate the distinct expertise of government, industry, and civil society actors (Rahimi and Jones, 2025). Establish joint response teams for space-cyber crisis management.
5. Enhance Cross-Sectoral Awareness and Resilience: Allocate resources for simulation-based awareness programs, institutional cybersecurity education, and psychological resilience initiatives. Encourage organizational risk assessment, proactive training, and the use of behavioral AI to detect emerging attack vectors, especially in resource-constrained settings.

7. Theoretical Frameworks and Models

7.1 Hybrid Warfare Space-Cyber Risk Assessment Framework

This paper proposes an expanded Space-Cyber Hybrid Attack Matrix, designed to meet contemporary standards for analytic rigor, incorporating multidomain vectors (cyber, space, information, and combined) that target critical infrastructure, social trust, command and control systems, the economy, and civilian populations through mechanisms such as malware, denial-of-service attacks, deepfakes, satellite jamming, algorithmic amplification, and psychological operations.

The Space-Cyber Hybrid Attack Matrix conceptualizes hybrid operations along three primary dimensions: vector, target, and mechanism. Table 1 presents example matrix cells that illustrate how these dimensions interact in practice. Vectors capture the dominant operational domain (cyber, space, information, or combined). Targets denote the principal object of harm (infrastructure, command-and-control, civilian population, social trust, or economy). Mechanisms describe the operational technique employed (e.g., malware, jamming, deepfakes, algorithmic amplification, or data exfiltration).

Table 1: Example matrix cells

Vector	Target	Mechanisms (examples)
Cyber	Infrastructure	Malware, ransomware, DDoS, data exfiltration
Information	Social Trust	Deepfakes, coordinated disinformation, bot amplification
Space	C2 / infrastructure	Satellite jamming, uplink spoofing, GNSS manipulation
Combined	Civilians, economy	Social-media-driven panic + satellite disruption of logistics

Analysts can use the matrix in scenario planning by mapping observed or hypothesized incidents into specific vector–target–mechanism cells. This enables comparison across cases, identification of dominant patterns (e.g., repeated attacks on social trust via information vectors), and prioritization of controls and policies for the most frequently exploited cells.

Organizations can use this model in scenario planning, contagion forecasting, and resource allocation for rapid response. Recent deployments in defense R&D labs indicate that continuous red-teaming, live incident simulation, and multi-layer risk analysis reduce overall exposure and improve incident response time by 18–32% (Mulahuwaish et al., 2025).

7.2 Algorithmic Amplification and Adversarial Operations

Algorithmic amplification models incorporating recommender engine dynamics, trending triggers, and multi-language campaign distribution now underpin both offensive and defensive operations. Advanced modeling demonstrates that content virality can be manipulated by strategic adversaries, requiring platform

operators to develop circuit-breakers and real-time feedback loops for rapid deplatforming or content suppression (Pennycook et al., 2021).

8. Limitations

Several limitations arise from the exclusive reliance on document-based analysis. The absence of direct interviews or practitioner testimony restricts the study's capacity to capture experiential, real-time insights into rapidly evolving threat landscapes and defensive practices. While secondary sources offer broad coverage, they may not fully reflect ground-level perceptions or implicit institutional challenges.

Second, data availability varied across regions, with certain states characterized by higher censorship, reduced incident disclosure, or incomplete access to technical documentation. Non-English and regionally disseminated sources may have been underrepresented despite efforts to ensure inclusivity. Temporal limits inherent to publication cycles further constrain the contemporaneity of analyzed incidents: developments occurring after major case documentation periods may not be fully captured.

Third, reliance on secondary literature introduces potential publication and selection bias, despite systematic attempts at triangulation and source corroboration. The study's findings and recommendations should therefore be interpreted as indicative of broader trends rather than definitive accounts of all operational realities.

Future research will benefit from integrating direct expert engagement, field interviews, and cross-linguistic content analysis to enrich contextual understanding and to validate emergent models under dynamic, multi-domain conditions. Expanding access to event-level data and increasing regional diversity in source material remain vital areas for methodological advancement.

9. Future Directions and Research Gaps

Quantum computing endangers conventional encryption standards for satellite and terrestrial communication, driving urgent research in post-quantum cryptography, multi-factor authentication, and secure key management.

AI-based defenders capable of learning from incident response patterns, adapting in real time, and cross-correlating behavioral cues hold considerable potential for next-generation resilience. However, their effectiveness in dynamic, multi-domain environments remains a subject for further empirical investigation.

Interdisciplinary collaborations across cyber psychology, technical operations, international law, and public health are increasingly critical. Replication studies and dynamic simulation exercises must be scaled to align with the evolving threat landscape. Further research must address the measurement of institutional and societal resilience, the operationalization of multi-domain deterrence, and the integration of next-gen digital forensics in both terrestrial and orbital contexts.

10. Conclusion

The escalating integration of social media, cyber operations, and satellite infrastructure creates a new paradigm for global security and hybrid warfare. As adversaries innovate, rapidly shifting tactics across physical, digital, and psychological domains, defenders must prioritize agility, intersectoral coordination, and robust technical standards. This paper synthesizes case analyses, published expert testimony, and cross-domain models to offer actionable policy recommendations and analytic frameworks. Resilience in the digital orbital era requires not only technical adaptation but legal harmonization, educational investment, and a commitment to global multi-stakeholder collaboration. The imperative to align international standards and pursue integrated multi-domain defense grows more pressing as attack vectors and vulnerabilities multiply across interconnected ecosystems.

Ethics Declaration: No human subjects, personal data, or confidential information were involved in this research, and clearance was not required.

AI Declaration: Generative AI assisted with wording and structure, while the author developed ideas, reviewed outputs, and ensured accuracy and integrity.

References

Bendett, S. (2023) 'Russia's hybrid warfare doctrine: Space, cyber, and information', Strategic Studies Review, 14(1), pp. 45–67.

Brandqvist, J. (2024) The cybersecurity threat of deepfake.

CIGI (2015) Hybrid warfare: Ukraine, Russia and Western lessons. Waterloo, ON: Centre for International Governance Innovation.

Clark, R. (2021) 'Cross-domain synergy and hybrid threats in modern conflict', Space Policy Quarterly, 17(4), pp. 239–254.

Cooley (2022) 'Key things to know about data protection laws in China', Cooley Privacy Talks. Available at: https://cdp.cooley.com/cooley-privacy-talks-key-things-to-know-about-data-protection-laws-in-china/ (Accessed: 27 November 2025).

CyberPeace Institute (2025) 'Cyber dimensions of a hybrid warfare: The case of Ukraine', CyberPeace Institute Reports. Available at: https://cyberpeaceinstitute.org/news/cyber-dimensions-of-a-hybrid-warfare/ (Accessed: 27 November 2025).

DataGuidance (2019) GDPR v. CSL and Specification: A comparative analysis. London: OneTrust DataGuidance. Available at: https://www.dataguidance.com/sites/default/files/gdpr_v_china_updated.pdf (Accessed: 27 November 2025).

Ferrara, E., Varol, O., Davis, C., Menczer, F. and Flammini, A. (2020) 'The rise of social bots and the attack on democracy', Communications of the ACM, 63(4), pp. 96–104. doi:10.1145/3377464.

Foust, J. (2019) 'The expansion of commercial space and security risks', Space News, 28(8), pp. 58–63.

Iskoujina, Z. and Roberts, T. (2024) 'Social media as an information warfare tool in the Russia–Ukraine conflict', Proceedings of the IDEAS Social Cybersecurity Conference. Carnegie Mellon University. Available at: https://www.cmu.edu/ideas-social-cybersecurity/events/ideas2024_paper_6.pdf (Accessed: 27 November 2025).

Li, Q. (2023) 'Influence of social bots in information warfare: A case study of the Russia–Ukraine conflict', Journal of Information Warfare, 22(2), pp. 45–67.

Marsh-Armstrong, B., Pacheco, F., Dameff, C. and Tully, J. (2024) 'Design and pilot study of a high-fidelity medical simulation of a hospital-wide cybersecurity attack', Research Square, rs-3.

Marwick, A. and Lewis, R. (2017) Media Manipulation and Disinformation Online. [online] Available at: https://datasociety.net/pubs/oh/DataAndSociety_MediaManipulationAndDisinformationOnline.pdf (Accessed: 27 November 2025).

Mozur, P. (2018) 'A genocide incited on Facebook, with posts from Myanmar's military', The New York Times, 15 October. [online] Available at: https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html (Accessed: 27 November 2025).

Mulahuwaish, A., Qolomany, B., Gyorick, K., Abdo, J.B., Aledhari, M., Qadir, J., Carley, K. and Al-Fuqaha, A. (2025) 'A survey of social cybersecurity: Techniques for attack detection, evaluations, challenges, and future prospects', Computers in Human Behavior Reports, 100668.

OECD (2023) Digital. [online] OECD. Available at: https://www.oecd.org/en/topics/digital.html (Accessed: 27 November 2025).

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G. and Rand, D.G. (2021) 'Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention', Frontiers in Psychology, 12, 646394. doi:10.3389/fpsyg.2021.646394.

Pillsbury (2021) China adopts new Data Security Law: Implications for cross-border data transfers. Pillsbury Winthrop Shaw Pittman LLP. Available at: https://www.pillsburylaw.com/en/news-and-insights/china-adopts-new-data-security-law.html (Accessed: 27 November 2025).

PwC (2021) 'How China's PIPL rules can impact your business', PwC Insights. Available at: https://www.pwc.com/us/en/tech-effect/cybersecurity/china-pipl-rules-impact.html (Accessed: 27 November 2025).

Rahimi, N. and Jones, H. (2025) 'Cyber warfare: Strategies, impacts, and future directions in the digital battlefield', Journal of Information Security, 16(2), pp. 252–269.

Tucker, J.A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D. and Nyhan, B. (2018) 'Social media, political polarization, and political disinformation: A review of the scientific literature', Political polarization, and political disinformation: A review of the scientific literature (19 March 2018).

Vavaroutsos, S. (2025) Modern warfare: The impact of social media on citizen journalism in interstate conflicts involving Ukraine. Master's thesis. Harvard University.

Secure Cross-Domain Data Validation with Field Programmable Gate Arrays

Abbie Cliche[1], James Brock[1], Ryan Longwell[1], Jason Dahlstrom[1] and Stephen Taylor[2]
[1]Web Sensing, LLC., Lebanon, USA
[2]Thayer School of Engineering at Dartmouth College, Hanover, USA

abbie.cliche@websensing.us
jim.brock@websensing.us
ryan.longwell@websensing.us
jason.dahlstrom@websensing.us
Stephen.taylor@dartmouth.edu

Abstract: Cross-domain systems have traditionally employed virtualization to isolate security domains, providing communication though standard TCP/IP networking stacks coupled with access permissions and credentials to enforce isolation. This deep and complex chain of trust typically depends upon a hardware base, such as a Trusted Platform Module (TPM) chip combined with a secure bootstrapping process. This paper describes a novel and high-performance alternative, Secure Transfer Link (STL), leveraging the unique architectural characteristics of the AMD UltraScale Multi-Processor System-on-Chip (MPSoC) device family: CPU affinity, an on-chip field programmable gate array (FPGA), and bus-mastering. These architectural characteristics make it possible to construct a secure data transfer path within the FPGA that can control which virtual machines may access and transfer data, enforcing isolation. The abstraction can be extended to include deep packet inspection and validation, such as parsing that checks adherence to the JavaScript Object Notation (JSON) protocol. Validation is achieved through the combination of formal grammars with a pushdown automata (PDA) parser and automatic transformation into an FPGA hardware configuration, resulting in a formally verifiable and hardened intellectual property (IP) called the Data Validator. The Secure Transfer Link is constructed by combining this Data Validator with another IP, the Memory Guard, which enforces access controls. These hardware Ips, together, comprise a system which prevents malicious software resident on a processor from undermining access policies or transferring malicious data. The presented IPs are performant. Their throughput improvement over traditional UDP/IP networking stacks is dramatic: speedups of up to 7x for tactical length messages and up to 4x for larger messages. The IPs are created using High-Level Synthesis (HLS), making it possible to formally specify a broad range of alternative policy and enforcement options then automatically include them in the Secure Transfer Link, constituting a novel isolation enforcement solution that is higher throughput than state of the art alternatives and enables selective domain access contingent upon formal verification of data.

Keywords: Hardware, Cybersecurity, Cross-domain, Data validation, Privacy, Secure data transfer, FPGA, SoC

1. Introduction

Ensuring the integrity, validity, and security of information transfers between security domains is becoming more important as combat arenas become saturated with high-value network-connected devices. This need is not unique to defense applications. Modern healthcare and critical infrastructure may be designed with separate privileged and public facing domains. A single device may be required to store information from multiple security classifications and incorporate information from a lower classification into higher classification domain to complete mission objectives. It is critical that the directionality of cross-domain transfers can be ensured, along with the credentials of the invoking process and the validity of the data itself. A typical approach to cross-domain transfers includes the Linux networking stack. However, when implemented in software, these mechanisms have been shown to be slow and cannot keep up with the needs of sophisticated real-time systems. Additionally, software is vulnerable to a number of API, buffer, and pointer attacks. Faster hardware-based mechanisms offer greater assurance but do not include robust payload filtering required to fully protect receiving domains. There remains a need for cross-domain data transfer inspection and validation technologies with high enough throughput and security assurance for critical systems. This paper presents two cross-domain transfer mechanisms. One is composed of custom parsing, validation, and authentication hardware IPs which are capable of performing deterministic transfer validation at throughputs of up to 7x traditional software alternatives. The other provides a network-based link through the FPGA fabric with the same validation IP. Prior work on automated hardware parser generation was leveraged in conjunction with the novel FPGA IP and cross-domain transfer methods presented in this paper to comprise the Secure Transfer Link.

2. Related Work

Many high value systems hold information at multiple classification levels and must enforce strict boundaries between them or require the ability to share information securely with stakeholders from different

organizations or nationalities (Dahlstrom *et al.*, 2019; National Security Agency (NSA), no date). It is vital that the direction of the flow of information can be guaranteed, that access policies can be enforced, and that data integrity can be guaranteed during these transfers. Cross-domain solutions are a class of technology which address this need, enabling controlled transfers of information between information domains with different security classifications or stakeholders (a *transfer* CDS) (General Dynamics Mission Systems, Inc, 2025; NSA, no date) or access to multiple information domains simultaneously (an *access* CDS) (Sundaravarathan *et al.*, 2024). An example of an access CDS is an application like SecureView (Dahlstrom *et al.*, 2019). According to the Air Force Research Lab (AFRL) (2021), SecureView is an application that allows a user to view multiple security domains in different windows on a single screen. Access and transfer cross-domain systems can either be a boon to productivity in enterprise environments or can provide critical inter-agency communication links in operational environments. Transfer CDSs can provide situational awareness through applications like a common operating picture (Sundaravarathan *et al.*, 2024). *Tactical* CDSs deployed in these environments require particularly low data transmission and processing latencies (General Dynamics Mission Systems, Inc, 2025).

A prototypical access CDS architecture may consist of a hypervisor running guest OSs connected by a network, coupled with a secure boot process, and rigorous mandatory access controls (MACs) (Daughety *et al.*, 2021; BAE Systems, 2022; AFRL, 2021; BAE Systems, no date; Everfox (2024)). A hardware root of trust could take the form of a TPM, a specialized cryptographic processor which can ensure the integrity of software used during the boot process (Pamnani and Matarazzo, 2024). Hypervisors are well established as a mechanism for providing isolation between guest OSs (Goldberg, 1974; Barham *et al.*, 2003; Thyagaturu *et al.*, 2022) on a single host and therefore can protect two domains collocated on a single chip. Doing so can reduce the size, weight, and power (SWaP) and latency of a system compared to an alternative architecture requiring physically separate machines. For access CDS applications in tactical domains, minimizing these metrics is always beneficial.

Successful domain isolation using virtual machines relies on privilege-based resource access and sharing. Hypervisors protect resources like memory ensuring that one application cannot surreptitiously gain access to another's memory space (Sundaravarathan *et al.*, 2024). Hypervisor-based CDSs are typically built on Type 1 hypervisors which behave like bare metal to the operating systems they support (AFRL, 2013; AFRL, 2021; Dahlstrom *et al.*, 2019; Sundaravarathan *et al.*, 2024). Formally proven secure hypervisors, such as the BlueRock Hypervisor (BHV) can provide additional assurance (Bedrock Systems Inc. (BSI), 2022).

Most hypervisors allow guest OSs to communicate with each other via either a virtual network or a shared memory mechanism. However, virtual networks do not typically come with data inspection and validation out of the box (Ren *et al.*, 2016; Green Hills Software, 2020; BSI, 2022). Use of network functions, in generals, is computationally expensive(Ren *et al.*, 2016; Thyagaturu *et al.*, 2022). Additionally, network-dependent CDSs like AFRL SecureView are vulnerable to network-based attacks from external interfaces (Sundaravarathan *et al.*, 2024). A single compromised virtual machine can compromise the other virtual machines on the host through either privilege abuse or steganographic attacks embedded in virtual network packets that pass all IP, port, and MAC checks (The MITRE Corporation, 2018). On the other hand, shared memory links between virtual machines can reach throughput exceeding 5Gbps but are highly coupled to a platform's fixed silicon architecture (Ren *et al.*, 2016; Li *et al.*, 2023). This means that there are no opportunities to insert security features, such as filtering, into the shared memory datapath.

A transfer CDS enables unidirectional or bidirectional secure transfers between domains. It may be implemented as software or hardware module (Sundaravarathan *et al.*, 2024) and may, like an access CDS, rely on a hardware base of trust (Daughety *et al.*, 2021). Transferring high bandwidth data from embedded sensors in tactical arenas is a key challenge in transfer CDS design because it requires performing filtering functions, which can be computationally expensive, at line rate (Dahlstrom *et al.*, 2019; Sundaravarathan *et al.*, 2024). Recently, transfer CDS architectures have favored filtering in hardware over filtering in software as the NSA has formally recognized hardware as part of robust system design in its Raise the Bar mandate (Campbell, 2019).

Syntactic filtering capabilities are a critical component of transfer CDSs. Syntactic filtering can stop attacks in which a file's contents don't match its extension and thwart attacks which depend on hiding malicious data between valid fields. When applied to data in transit, filtering can protect against vulnerable APIs in receiving software that do not adequately validate data (The MITRE Corporation, 2020). The challenge historically has been accomplishing syntactic filtering on data payloads with low latency (Swarup, 1994). Consequently, payload parsing is sometimes executed as post-processing or is only applied to a subset of packets (Hu *et al.*, 2023).

There are several hardware-enforced cross-domain solutions on the market (BAE Systems, 2020; "XTS® Diode One Way Transfer Solution," 2022; Everfox, 2024). Some claim to reach 10Gbps of throughput ("XTS® Diode One

Way Transfer Solution," 2022), but this value reflects the maximum throughput under laboratory conditions, and it is unlikely that it includes content filtering. Others offer semantic data inspection and sanitization but do not advertise throughput (Everfox, 2024). One experimental application of hardware filtering reaches 3Gbps, but it employs payload inspection only on packets that have failed header inspection, was only evaluated on packets up to 512 bytes in length, and did not document clock speeds used in the FPGA fabric (Hu *et al.*, 2023). Several software-based systems, such as a secure RTOS separation kernel, do not provide content filtering for transfers between domains (Green Hills Software, 2020; Daughety *et al.*, 2021). Even if content filtering were included, it would still be slower and more power intensive than equivalent hardware-based content filtering. As little as 3Mbps of data throughput would be sufficient for transmission of video, so a hardware-based syntactical filter with performance exceeding that threshold, like the work we present here, would constitute a step forward in the state of the art.

CPU affinity is a guest VM setting supported by most hypervisors. VMs can be locked to a particular CPU core by the hypervisor and therefore predictably linked to other hardware resources on an SoC with virtualization support (Dahlstrom *et al.*, 2019; "Zynq UltraScale+ Device Technical Reference Manual," 2023). In the case of the UltraScale+ MPSoC, this is accomplished through the use of AXI side channels, TrustZone technology, and the XMPU (McNeil *et al.*, 2021). These technologies enable VM resource isolation which extends to the FPGA fabric, forming a trusted connection between a VM and IP blocks residing in the FPGA fabric. Together, these mechanisms make it possible to create a shared memory between virtual machines with hardware-enforced access controls. The speed and flexibility of the FPGA fabric open up new options for adding deep packet inspection and validation capabilities to a system. Formally defined parsers can be constructed quickly and easily using Vitis HLS and then deployed to provide deterministic data filtering (Taylor *et al.*, 2022). Even when validation processes add latency to a system, those run in FPGA fabric will outperform their software counterparts.

3. Methods

We have established the general notion that routing cross-domain transfers through the FPGA fabric could lead to improved performance over the network stack. By timing data transfers between two Linux virtual machines (VM0 and VM1) running on the BlueRock Hypervisor, we collected two sets of data to establish a baseline to compare our solution against. The first is the latency of transferring data through BlueRock's vSwitch and the second is the latency of transfers routed through a hard-wired 1Gbps ethernet link through the FPGA fabric, referred to as the *fixed link* from here on out. We then measure the latency of data transfers on our novel cross-domain transfer mechanism, known as the *Secure Transfer Link (STL)* from this point forward. The STL comprises custom memory-mapped IPs in the FPGA fabric responsible for gating each virtual machine's access to memory regions by checking user permissions and validating transfers between memory regions with a data payload validator. Figure 1 shows the three different data transfer pathways within the FPGA SoC.

All of our experiments were performed on the Zynq UltraScale+ MPSoC ZCU102 Evaluation Kit using a programmable logic (PL) fabric clock speed of 100MHz and 1G Ethernet links (shown in by the gigabit media independent interfaces [GMII] in Figure 1). The test data used in the experiments was a set of text files containing JavaScript Object Notation (JSON) structures generated by a fuzzer. The set contained a uniform distribution of structures with depths ranging from 1 to 15.

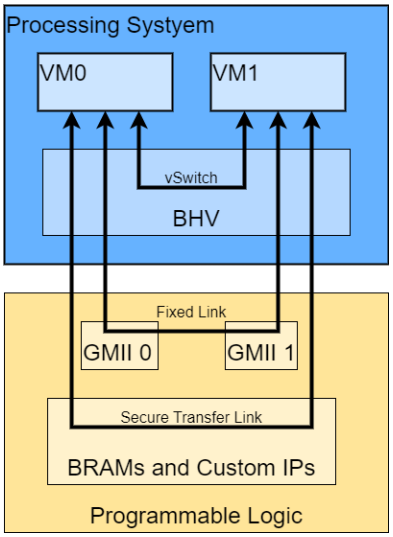


Figure 1: The three cross-domain transfer mechanisms tested in this paper

3.1 Hardware Design

This work includes two encapsulations of previous a previously developed automatic grammar parser and validator generation flow (Taylor *et al.*, 2022; Taylor and Pope, 2024). This parser was constructed by transforming a formal JSON grammar definition into parsing tables which were compressed and inserted into a generic C++ push-down automata (PDA) implementation. Vitis HLS was leveraged to automatically generate a grammar validator hardware IP module from this source code. This process is generic, and validators for other grammars can be constructed from a number of supported grammar tables which are available as an open-source software library (Taylor *et al.*, no date). The PDA can currently reach a JSON validation throughput of 24Mbps (Taylor *et al.*, 2022).

The fixed link shown in Figure 1 connects two of the MPSoC's integrated Ethernet blocks in a bidirectional direct 1Gbps configuration (VM0 to VM1 and VM1 to VM0). The fixed link also has two modes, *passthrough* and *validation*. In the passthrough mode, the Ethernet link contains no additional logic. In the validation configuration, the Ethernet link contains an instance of our PDA parser encapsulated in a network packet parser. The network packet parser validates the IP, Ethernet, and TCP/UDP packet headers and then extracts the JSON payload for the PDA to validate. The encapsulating IP can be configured in validation mode, when packets whose payloads do not pass the grammar check are dropped, or in passthrough mode, in which all packets pass through the link.

The Memory Guard IP is a build-time configurable hardware block capable of enforcing complex access rules on a variety of memory mapped regions held within the FPGA PL fabric. The Memory Guard can separate the memory in its address map into many segments defined by the user. Read/write permissions can be granted per memory segment and per originating CPU core. Because the VMs in our experiments are pinned to specific CPU cores, permissions can be set based on the privilege level of the VM. Enforcement of privileges is accomplished by observing AXI side-channel signals to detect which CPU core initiated each memory transaction. There is an additional reserved "honeypot" memory region, which can be enabled and configured. Memory transactions which fail permissions checks can be diverted to this region so that they appear to have succeeded but don't allow unauthorized access to data. The Memory Guard can be placed on any memory mapped processing system interface, enabling programmable access controls and responses at any security barrier in the system.

The data validator (DV) is the hardware module developed to validate payload data in flight as it is transferred between memory regions in two modes: validation and passthrough. The Data Validator IP has two 32-bit wide AXI memory mapped interfaces, each of which is connected to a BRAM (block random access memory) within the FPGA fabric. The data validator reads data from the *origin* BRAM and passes it into the PDA JSON parser (if validation mode is enabled). If the JSON parser check fails at any point, the data is dropped and the BRAMs cleared. Otherwise, the data is written into the second *destination* BRAM. In passthrough mode, all data is transferred to the destination BRAM. The Data Validator and processing system use mailbox bits for synchronization. If data is dropped, a status indicator is raised and can be consumed by software as an interrupt so that application level logic can handle retransmission of messages if applicable for a given use case.

The Zynq UltraScale+ MPSoC ZCU102 Evaluation board contains an SoC with a full quad-core ARM A53 processing system (PS) co-located with the programmable logic (PL) resources of a large FPGA. The large number of available direct hardware interfaces between the processor and this FPGA include Ethernet and AXI memory mapped interfaces. The Secure Transfer Link shown in Figure 1 utilizes the AXI memory mapped interfaces for mapping PL BRAMs to the VMs in the processing system. The Secure Transfer Link itself is composed of two individual data diode links configured in opposite directions, one transferring data from VM0 to VM1 and the other transferring data from VM1 to VM0, as shown in Figure 2. Each link consists of two BRAM blocks connected by an instance of the Data Validator IP, and PS access to each BRAM is gated by an instance of the Memory Guard IP.

Our experimental platform is meant to represent a system with multiple domains of varying levels of privilege. The two VMs isolated by the BlueRock Hypervisor represent those domains. The shared memory blocks are first protected by our Memory Guard (MG) IP with permissions similar to the shared memory region in Figure 2. The difference in our Secure Transfer Link test setup in Figure 2 is that we have shared memory regions moving data bidirectionally through two separate links to use the round-trip timing in our results.

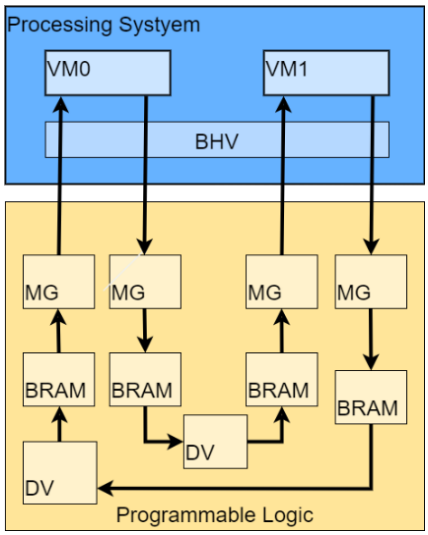


Figure 2: The secure transfer link

3.2 Test Applications

VM0 and VM1 run on top of the BlueRock Hypervisor (BHV). We used the BHV's built-in virtual switch (vSwitch) as a pure software transfer link to compare against our FPGA fabric mechanisms. To test either the vSwitch or the fixed link, a network client program on VM0 sends a UDP packet out over the relevant interface. The UDP packet is received and sent back by the server program to the original source VM to generate a complete round-trip time for data transfers. When measuring the fixed link, the elapsed time includes overhead incurred by the Linux network stack as well as overhead from the Ethernet frame, IP and UDP headers, and the interframe gap. When measuring the vSwitch mechanism, overhead is only incurred by virtual switching.

A separate application was used to test the Secure Transfer Link. It drives raw data to the Memory Guard, which drives the data onward through the rest of the Secure Transfer Link. The corresponding application running on VM1 polls its shared memory for data and reassembles the received data as a txt file. It then writes this file back onto the Secure Transfer Link in the other direction. The test program running on the source VM collects the elapsed time between the first chunk of JSON data being written to the origin shared memory and the last chunk being read from the destination shared memory.

We computed latency and throughput from the measured elapsed time of a full return trip. We ran tests transferring data sizes from 1 to 1472 bytes to gauge the performance of each transfer link as the transfer size increases. Each test was repeated 10 times to get an average performance and account for random software anomalies in any given run. The latency data presented is the measured latency divided by two to compensate for the inclusion of a return trip. Throughput was computed by dividing the number of bits in each test file by the latency of the transfer.

Latency measurements were taken for the following data transfer mechanisms: Secure Transfer Link passthrough mode, Secure Transfer Link validation mode, fixed link passthrough mode, fixed link validation

mode, and BHV vSwitch. The VSwitch does not have any payload parsing capabilities, so it is only configured as a passthrough.

Each of the JSON data files used in these experiments was generated using a fuzzer. The files ranged in size from 1 to 1472 bytes and had random key-value pairs and nested structures. The maximum JSON structure depth used in these experiments was 15 because it was impractical to create deeper structures in the smallest files. Only valid JSON structures were used for performance testing since invalid structures would introduce non-deterministic latency decreases as the result of halting processing after reaching an error in the data structure and not from any advantage of the secure transfer link.


```
{
    "LabIV8E": -2006,
    "q3": false,
    "1CP": 724.8507268121102,
    "cOPY4K": "VwxslZatE",
    "2Mfz6": 424
}
```


Figure 3: Sample JSON structure with a depth of 1 used in the experiment

Randomized raw JSON, as shown in Figure 3, was used for the Secure Transfer Link tests, but it was encapsulated in a UDP packet for tests on the Fixed Link and vSwitch link. UDP was selected over TCP to minimize non-payload overhead as much as possible. A maximum payload size of 1472 was used because common tactical and "command and control" traffic using protocols such as MAVLink fit within packets of this size, and it is the maximum data size for a single UDP packet (Taylor *et al.*, 2022).

4. Results

The bandwidth graph in Figure 4 compares the performance of the BlueRock vSwitch data transfer mechanism with that of our fixed link passthrough configuration. The measurements of the two mechanisms are nearly identical, with a root mean squared error of 60µs between the two sets of latency measurements and a root mean squared error of 23Kbps between the two sets of throughputs. Through this comparison, we have established the feasibility of routing data transfers through the FPGA fabric and shown that there is not a significant overhead or slowdown incurred from doing so. We have shown that including the FPGA fabric in the data path enables packet validation logic is enabled at no cost.

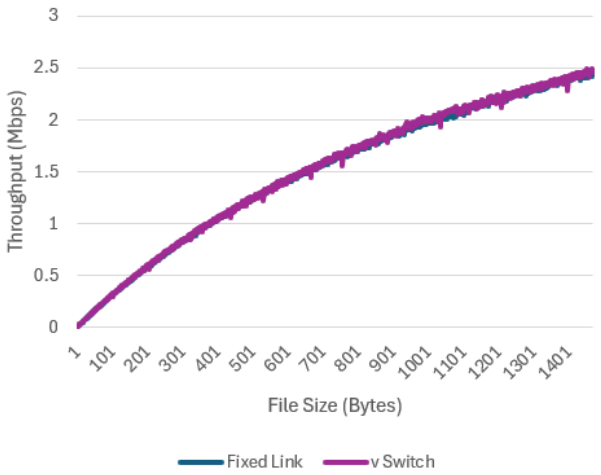


Figure 4: Fixed link vs. vSwitch passthrough bandwidth

We now compare the Secure Transfer Link performance against the hardware fixed link, which we have established to have similar performance to the BHV vSwitch method of cross-domain transfer. We first compare the two links with no parsing inline to establish baseline performance improvement. Latency of the fixed link is greater than 2ns and up to 3.5ns more than the latency of the Secure Transfer Link and twice as responsive as that of the Secure Transfer Link to the size of the JSON data. This shows that cross-domain transfers of tactical length messages over the Secure Transfer Link can be performed with about a quarter of the latency by bypassing the overhead of the network stack.

Figure 5 shows the bandwidth of same set of tests. The bandwidth of the fixed link remains low for all sizes of JSON message due to the significant overhead involved in the software network stack. Tactical messages are not large enough relative to the overhead to make the fixed link an efficient method of cross-domain data transfer. In contrast, the bandwidth of the Secure Transfer Link increases as the JSON data size grows due to the small latency overhead of the hardware mailbox signaling and low-latency data movement. This figure shows that bypassing the network stack using the Secure Transfer Link leads to up to 7x bandwidth improvement for cross-domain transfers. A novel shared memory mechanism such as our Secure Transfer Link can yield greater throughput and lower latency when the message size is small enough that the network stack overhead would not be amortized away.

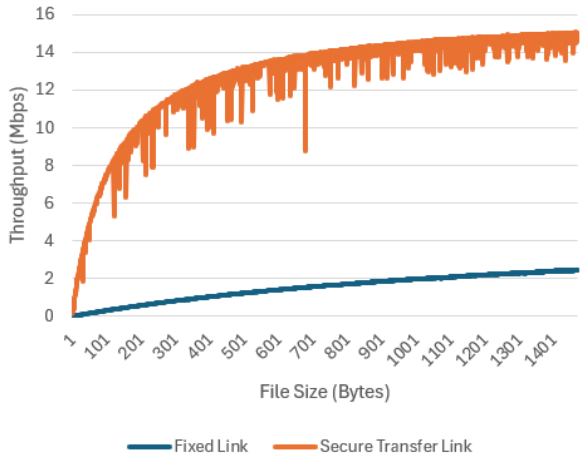


Figure 5: Fixed link vs. secure transfer link passthrough bandwidth

Now we introduce the Data Validator with formal JSON parsing into the Secure Transfer Link data path. The latency resulting from including JSON validation in the Secure Transfer Link is greater than the latency without for all data sizes and increases linearly with data size. However, the maximum latency increase for tactical messages is only 407µs. Figure 6 shows that the cost of validation with including the JSON validation is, at worst, a 28% decrease in bandwidth. The resulting 10Mbps throughput is still more than sufficient for applications like streaming video calls and well exceeds the performance required of a tactical CDS.

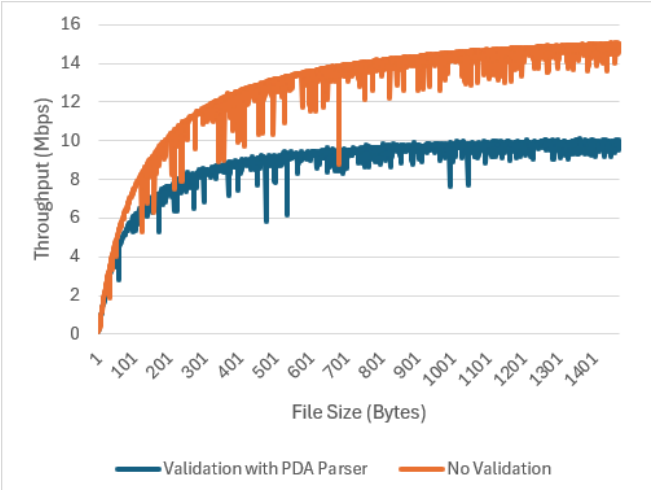


Figure 6: Secure transfer link passthrough vs. secure transfer link validation bandwidth

5. Conclusion

The number of cross-domain systems employed in modern tactical environments is growing, and performance and SWaP (size, weight, and power) constraints are tightening. This research demonstrates a novel combination of cross-domain assurance through syntactical data validation and a high throughput hardware channel. Our Secure Transfer Link can protect against data injection attacks through deterministic grammar checkers and guard against memory operations initiated by virtual machines in the system which are not whitelisted,

preventing information from leaking between security domains. Additionally, automatically generating FPGA IP from C++ source enables this solution to be easily tailored to other grammars and deployed to extremely low-SWaP environments without FPGA design expertise.

We showed that routing cross-domain transfers through the fabric, in general, is at least as performant as routing through a hypervisor's virtual switch by comparing our fixed-link mechanism to the BHV VSwitch. We have also demonstrated that the Secure Transfer Link attains as much as a 7x speedup over links that include the network stack for tactical length messages. Finally, we showed that data validation mechanisms resident in the FPGA fabric provide assurance without undermining the performance gains of our shared memory mechanism by demonstrating a throughput of up to 10Mbps. This performance beats all other cross-domain links demonstrated in this paper, including the vSwitch which does not include data validation. It also beats the throughput of state-of-the-art software-base syntactic parsing demonstrated in related work (Hu *et al.*, 2023).

Acknowledgements

This research is supported by the Defense Advanced Research Projects Agency (DARPA).

DISTRIBUTION A : Approved for Public Release; Distribution is unlimited.

Ethics Declaration: Ethical clearance was not required for the research carried out in this paper.

AI Declaration: AI tools were not used in producing this paper.

References

Air Force Research Laboratory (2013) "SecureView Overview", Air Force Research Laboratory. Available at: https://www.ainfosec.com/wp-content/uploads/2015/09/SecureView_Overview_Master_PA_Cleared_7Oct13.pdf.

Air Force Research Lab (2021) "AFRL_SecureView_FS_0922", Air Force Research Lab. Available at: https://afresearchlab.com/wp-content/uploads/2021/04/AFRL_SecureView_FS_0922.pdf (Accessed: November 5, 2024).

BAE Systems (no date) "STOP™ High Assurance GPOS" *BAE Systems*. Available at: https://www.baesystems.com/en/product/stop-os (Accessed: November 9, 2024).

BAE Systems (2020) "Data Diode Maximum assurance for unidirectional throughput", BAE Systems.

BAE Systems (2022) "XTS® Diode One Way Transfer Solution", BAE Systems. Available at: https://www.baesystems.com/en-media/uploadFile/20230504165402/1434650747473.pdf.

Barham, P. *et al.* (2003) "Xen and the art of virtualization," in *Proceedings of the nineteenth ACM symposium on Operating systems principles*. *SOSP03: ACM Symposium on Operating Systems Principles*, Bolton Landing NY USA: ACM, pp. 164–177. Available at: https://doi.org/10.1145/945445.945462.

Bedrock Systems Inc. (2022) "Bedrock Hypervisor (BHV) User Guide" (2022), BedRock Systems Inc.

Campbell, S. (2019) "A Brief Note on Raise the Bar and One-Way Transfer," *OWL Cyber Defense*, 18 September. Available at: https://owlcyberdefense.com/blog/a-brief-note-on-raise-the-bar-and-one-way-transfer/ (Accessed: January 15, 2025).

Dahlstrom, J. *et al.* (2019) "Hardening Containers for Cross-Domain Applications," in *MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM)*. *MILCOM 2019 - 2019 IEEE Military Communications Conference (MILCOM)*, Norfolk, VA, USA: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/MILCOM47813.2019.9020992.

Daughety, N. *et al.* (2021) "vCDS: A Virtualized Cross Domain Solution Architecture," in *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*. *MILCOM 2021 - 2021 IEEE Military Communications Conference (MILCOM)*, San Diego, CA, USA: IEEE, pp. 61–68. Available at: https://doi.org/10.1109/MILCOM52596.2021.9652903.

Everfox (2024) "Datasheet High Speed Verifier (HSV)" (no date), Everfox. Available at: https://cms-assets.everfox.com/app/uploads/2024/03/19161730/Datasheet_High-_Speed_Verifier_HSV-en-021324.pdf.

General Dynamics Mission Systems, Inc (2025) "Frequently Asked Questions For Cross Domain Solutions", *General Dynamics Mission Systems*. Available at: https://gdmissionsystems.com/cross-domain-solutions/faqs (Accessed: January 12, 2024).

Goldberg, R.P. (1974) "A survey of virtual machine research," *Computer*, 7(6), pp. 34–45.

Green Hills Software (2020) "Creating a Trusted Embedded Platform for MLS Applications", Green Hills Software. Available at: https://www.ghs.com/download/whitepapers/GHS_Trusted_Platform_MLS.pdf.

Hu, Z. *et al.* (2023) "Realtime Malicious Traffic Detection Targeted for TCP Out-of-Order Packets Based on FPGA," *IEEE Access*, 11, pp. 112212–112222. Available at: https://doi.org/10.1109/ACCESS.2023.3323853.

Li, Z. *et al.* (2023) "RTISM: Real-Time Inter-VM Communication Based on Shared Memory for Mixed-Criticality Flows," in *2023 IEEE Real-Time Systems Symposium (RTSS)*. *2023 IEEE Real-Time Systems Symposium (RTSS)*, Taipei, Taiwan: IEEE, pp. 279–290. Available at: https://doi.org/10.1109/RTSS59052.2023.00032.

McNeil, S. *et al.* (2021) "Isolation Methods in Zynq UltraScale+ MPSoCs."

The MITRE Corporation (2018) "CAPEC-122: Privilege Abuse." Available at: https://capec.mitre.org/data/definitions/122.html (Accessed: January 28, 2025).

The MITRE Corporation (2020) "CAPEC-113: Interface Manipulation." Available at: https://capec.mitre.org/data/definitions/113.html (Accessed: January 28, 2025).

National Security Agency (no date) "National Cross Domain Strategy & Management Office", *National Security Agency/Central Security Service*. Available at: https://www.nsa.gov/Cybersecurity/Partnership/National-Cross-Domain-Strategy-Management-Office/ (Accessed: November 9, 2024).

Pamnani, V. and Matarazzo, P. (2024) "Trusted Platform Module Technology Overview," *Microsoft Learn*, 10 July. Available at: https://learn.microsoft.com/en-us/windows/security/hardware-security/tpm/trusted-platform-module-overview (Accessed: November 2, 2024).

Ren, Y. *et al.* (2016) "Shared-Memory Optimizations for Inter-Virtual-Machine Communication," *ACM Computing Surveys*, 48(4), pp. 1–42. Available at: https://doi.org/10.1145/2847562.

Sundaravarathan, V. *et al.* (2024) "Cross-Domain Solutions (CDS): A Comprehensive Survey," *IEEE Access*, pp. 1–1. Available at: https://doi.org/10.1109/ACCESS.2024.3483659.

Swarup, V. (1994) "Automatic Generation of High Assurance Security Guard Filters", *Proceedings of the 17th National Computer Security Conference (NCSC*, Volume 1), pp. 123–131.

Taylor, S. *et al.* (2022) "Automatic Construction of Hardware Traffic Validators," *European Conference on Cyber Warfare and Security*, 21(1), pp. 60–69. Available at: https://doi.org/10.34190/eccws.21.1.200.

Taylor, S. *et al.* (no date) "thayer_parsers." Available at: https://github.com/lvln/thayer_parsers (Accessed: February 10, 2025).

Taylor, S. and Pope, G. (2024) "Hardware Sequence Combinators," *International Conference on Cyber Warfare and Security*, 19(1), pp. 358–365. Available at: https://doi.org/10.34190/iccws.19.1.1965.

Thyagaturu, A.S. *et al.* (2022) "Operating Systems and Hypervisors for Network Functions: A Survey of Enabling Technologies and Research Studies," *IEEE Access*, 10, pp. 79825–79873. Available at: https://doi.org/10.1109/ACCESS.2022.3194913.

"Zynq UltraScale+ Device Technical Reference Manual" (2023).

Malinformation, Deepfakes, and Cyber Warfare: Ethical and Anticipated Ethical Issues

Noah Donnelly[1] and Richard Wilson[1,2]
[1]Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA
[2]Department of Philosophy, Towson University, Baltimore, Maryland, USA

ndonnel1@students.towson.edu
wilson@towson.edu

Abstract: Malinformation and deepfakes are closely connected in cyber warfare because both involve the use of *true* or *realistic-looking* information in harmful ways. Malinformation is *genuine (truthful) information used maliciously* to cause harm. Unlike disinformation (false) or misinformation (false but unintentional), malinformation is based on truth, but truth that has now been weaponized. Examples include, leaking private emails, exposing sensitive military data, or releasing authentic documents but out of context. Deepfakes can be AI-generated synthetic media (video, audio, images) that convincingly mimic real people or events. They can be employed to fabricate speeches, fake evidence, or impersonate leaders. Malinformation and Deepfakes Intertwine in several ways. They involve authenticity exploitation. Deepfakes often mix real content with fabricated elements. When genuine material is combined with manipulated media, it becomes malinformation — because the *truthful parts* lend credibility to the fake narrative. *Example:* Real leaked emails paired with a deepfake video of a politician "admitting" corruption. They promote context manipulation. Malinformation thrives on taking real information out of context. Deepfakes amplify this by visually or audibly presenting "evidence" that seems authentic. *Example:* A real battlefield video edited with deepfake audio to suggest atrocities that didn't occur. They have a psychological impact. Malinformation already erodes trust by exposing sensitive truths. Deepfakes magnify this by making it harder to distinguish between genuine leaks and fabricated ones. *Example:* Soldiers' real identities leaked (malinformation) alongside deep-fake videos of them committing crimes. Cyber warfare Strategy. Both are used to destabilize societies, discredit leaders, and manipulate public opinion. Deepfakes transform malinformation into a more persuasive weapon by adding *visual proof*. Analysts' warn that the "malinformation–deepfake nexus" is one of the most dangerous aspects of modern cyber warfare. Malinformation and deepfakes become tools of cyber warfare by bringing truth and illusion together, making it extremely difficult for societies to defend against manipulation. Malinformation and deepfakes intersect with cyber warfare because both weaponize elements of truth and realism to cause harm, erode trust, and destabilize societies. Deepfakes often transform malinformation into a more persuasive and damaging tool. This analysis will identify the ethical and anticipated ethical issues with the use Malinformation and Deep Fakes in Cyber Warfare.

Keywords: Malinformation, Deepfakes, Generative adversarial networks (GANs), Epistemic backstop, Liar's dividend, C2PA, Information disorder

1. Introduction

Within the conflicts of the modern era, the battlefield has grown past kinetic destruction of infrastructure to an additional focus on the cognitive alteration and destruction of reality itself. Wardle and Derakshan (2017) categorize this as "Information Disorder," which is characterized by three unique traits: Misinformation (unintentional lies), Disinformation (intentional lies), and Malinformation (weaponized truth). Cyber warfare traditionally focuses on the disruption of networks, where the new frontier is on the corruption of data. Two technologies have been the primary offenders in terms of this corruption: Malinformation, which uses breached private data to destroy people and institutions reputations, and Deepfakes, which uses Generative AI to create fake evidence.

The combination of Deepfakes and Malinformation creates a "Liar's Dividend," an idea introduced by Chesney and Citron (2019). The Liar's Dividend posits that when high-quality fake media becomes more prevalent, rational actors will begin to discount all media, media that includes organic evidence of corruption or war crimes. A state actor now has the capability of leaking real compromising files (Malinformation) and of combining them with deep-fake videos that create a fake but incriminating context. The point of these actions is not just to fool people, but to pollute the information pool so much that the targeted population begins to disengage from politics entirely, this tactic is known as "censorship through noise." This paper analyzes how the technical mechanisms behind Generative Adversarial Networks (GANs) violate the ethical "Right to Truth," and applies Anticipatory Ethics to the rampant, unchecked proliferation of open-source synthetic media technologies. Through the cases of the Ukraine conflict, the Gabon coup attempt, and the Macron Leaks, we show that the target of modern information warfare operations is the human mind itself.

2. Technical Issues

Synthetic media is being weaponized, and this weaponization is being driven by the rapid commoditization of deep learning models. Early deepfakes relied on basic autoendcoders, neural networks capable of learning data coding's in an unsupervised way, the current threat is defined by three, very advanced technical artifacts: Latent Diffusion Models, Transformer-based Audio Synthesis, and Adversarial Counter-Forensics (Mirsky & Lee, 2021).

2.1 Latent Diffusion Models

General Adversarial Networks (GANs) originally sparked the deepfake era by the "minimaxing" game between Generator and Discriminator. The new method is Latent Diffusion Models (LDMs). GANs map noise directly to an image, LDMs are far more advanced, they operate by iteratively denoising a random Gaussian noise field based off text or semantic inputs (Rombach et al., 2022). This allows for more stability and a higher resolution then GANs have, which typically suffer from "mode collapse," where the generator would output an image identical to the one it was provided. In the context of cyber warfare, LDMs represent force multipliers, the feature of "Inpainting" (ability to edit specific regions of a real image) and "Outpainting" (ability to insert compromising elements like nudity, illicit substances), all done with perfect coherence. LDMs lower the technical barrier for high-fidelity forgeries. LDMs simply do not need the complex hyperparameter tuning that GANs require (Mirsky & Lee, 2021).

2.2 Zero-Shot Audio Synthesis and Transformers

Possibly the most frightening evolution is the rise of Zero-Shot Text-to-Speech (TTS). Previous models like Tacotron 2 needed hours of a targets voice to train a clone of a voice. Modern TTS's are built on Transformer models (similar to GPT-4) which treat audio synthesis as a language modeling task. By turning audio waveforms into acoustic tokens, models such as VALL-E can mimic the targets voice, prosody, and emotional tone in as little as three seconds of audio as a reference (Wang et al., 2023). Using this capability, attackers can simply intercept a phone call, train a clone, and disseminate falsities in a matter of minutes (Malinformation). The attacker would take a real, leaked phone call, use a Speech-to-Speech translator that retains the target's vocal signature and the correct background noise, which makes modern spectral analysis ineffective (Westerlund, 2019)

2.3 Adversarial Perturbation and Counter-Forensics

With the improvement of detection algorithms, attackers are adapting by deploying Adversarial Perturbations. This involves injecting a specific pattern of low-magnitude noise, invisible to humans, into the deepfake which mathematically blinds the detection algorithms (Carlini & Farid, 2020). An example of such tactics would be an attacker adding a gradient-based noise layer that would force a RestNet-50 detector to mark a fake video as "Real" with 99% certainty. This is the "Arms Race" that is dynamically unfolding on the digital stage; aggressors optimize their content to defeat the forensic tools that platforms like Facebook or Twitter employ. Passive detection is unreliable because the detector is now the target of the generative process.

2.4 Malinformation Infrastructure

The production of Malinformation is no longer a manual process, it is automated using Natural Language Processing (NLP) pipelines. When an Advanced Persistent Threat (APT) steals terabytes of data, attackers use Named Entity Recognition (NER) to automatically scan millions of files for sensitive keyworks like "confidential," "bank," "health," and "affair." The leaks are then indexed by the systems which now allow intelligence officers to query for compromising material related to specific targets. Now automated, it allows for leaks to be "Micro-Targeted" where subsets of private data are released to a target demographic on social media to maximize psychological impact. This is known as "Precision Propaganda" (Starbird et al., 2019).

3. Ethical Issues

Deploying these technologies in warfare raises important questions regarding the "Right to Truth."

3.1 Eroding the Epistemic Backstop

Philosopher Regina Rini (2020) argued that audiovisual recordings serve the purpose of an "Epistemic Backstop," which is a final, objective verifier of 'what is reality' that resolves disputes. When that verifier (video evidence) is undermined, truth itself becomes hard to ascertain and since the backstop is removed, society enters a state of "Epistemic Insecurity." This attacks the autonomy of the citizens themselves, which can be defined as an attack on a civilian population, which is an ethical violation of the laws of war. If citizens cannot determine if a video about a presidential candidate is real or not, they can no longer make a free and rational decision about

the reliability of the candidate. This causes the degradation of the democratic process and moves society towards tribal beliefs and not evidence-based good, spirited debate on important issues. Fallis (2020) expands upon this, arguing that the existence of the technology causes harm by creating "Technological Defeasibility," which is the action of disregarding true evidence believing it is fake. This allows authoritarian regimes to write off documented human rights abuses as "Western AI fabrications," protecting these regimes from accountability and the repercussions that come as the result of accountability.

3.2 Digital Violence and Contextual Integrity

"Contextual Integrity" proposed by Nissenbaum (2019) is an ethical framework violated by Malinformation. Malinformation violates the typical channels in which information flows, typically governed by context-specific norms (medical data staying with a doctor), is now taken out of context and weaponized by forcing private information out of its intended context, into the public realm. Even if the information is technically true, its release is intended to coerce or shame the target. In cyber warfare, this is targeting the "Human Layer" of the Open Systems Interconnection (OSI) model (Zimmerman, 1983). By exposing the private lives of important individuals like diplomats, politicians, and soldiers, the attackers aim at causing psychological trauma that can destroy their adversary's motivation to fight. This is a violation of the Just War principle of Proportionality (May, 2007), as the long-term effects upon the targets psyche represent damage to non-combatants (in a reputational sense), which far exceeds the military advantage gained through kinetic warfare.

4. Case Studies

Cyber criminals leverage these attacks to maximize economic chaos and instigate liquidity crises (Kenton, 2023). The once theoretical threat of deepfake-malinformation has already materialized in global conflicts, altering the course of history. These cases showcase how state and non-state actors use the "Liar's Dividend" and "Tainting" strategies to bypass traditional defense systems against deepfakes.

4.1 Zelenskyy Fake Surrender Video (2022)

In March of 2022, a hacker group compromised the systems of the Ukrainian news station called *Ukraine 24* by hijacking the live news feed to stream a deep faked video of President Volodymyr Zelenskyy. In this fake video, Zelenskyy stood behind a podium and ordered all Ukrainian troops to lay down their arms and surrender themselves to the Russian forces. Even though the quality of the video was low, it featured a static body and mismatched pixelation around Zelenskyy's neck, nevertheless the operation showed the concept of a "Deception Strike" via information warfare. The hackers didn't just rely on social media, they compromised a "root of trust," that being the national news broadcast, to attempt to lend credibility to the fake video. Atlantic Council analysts noted that while this attempt failed due to rapid debunking, it was considered a "Proof of Concept" for future operations. If such a video was released during a time of blackout or nuclear warfare, soldiers on the front lines might obey the order before it could be fact checked. This shows the dangers of deepfakes in "Time-Critical Disinformation" operations related to kinetic warfare (Atlantic Council, 2022).

4.2 Gabon Coup Attempt (2018)

In late 2018, the president of Gabon, Ali Bongo, had not been publicly seen for months due to a stroke, which made a power vacuum fueled by rumors of his death. To create an image of stability, the government released a "New Year's Address" video. The video was uncanny, observers noticed the president's eyes did not blink for long periods of time and his facial movements appeared to be like someone wearing a mask. The political opposition instantly claimed that the video was a deepfake, and proof that the president was either dead or incapacitated. Claiming that this "fake" video was authentic meant that the government had lost all legitimacy, As a result Lieutenant Kelly Ondo Obiang launched a military coup on January 7, 2019. This case is an example of the "Liar's Dividend": the existence of deepfake technology alone had allowed the rebels to weaponize speculation, transforming a low-quality video to support rumor and using the video as a justification for a violent military coup. This gave evidence that deepfakes do not need to be high quality to be lethal, they only need to be plausible enough to confirm an existing bias (Breland, 2019).

4.3 Macron Leaks (2017)

Hackers linked to the Russian GRU released 9 gigabytes of emails that were stolen from Emmanuel Macron's campaign in an operation called #MacronLeaks, two days before the French presidential election. This was a professional "Malinformation" operation that used a technique called "Tainting." The dump was primarily boring, authentic administrative emails, but the hackers forged a small number of documents suggesting that

Macron had offshore bank accounts containing illicit funds. The release was timed for Friday night before the election, triggering the mandated 44-hour media blackout in France. This timing was planned so that journalists could not verify the documents before the voting began, forcing the French public to rely on social media rumors. The operation only failed because the campaign had flooded their own servers with a “honeypot” of fake data which confused the attackers. Nevertheless, it established a doctrine of “Hack-and-Forge,” where truth is used to deliver lies (Jeangène Vilmer, 2019).

4.4 Pentagon Explosion Hoax (2023)

In May of 2023, a verified Twitter account by the name of “Bloomberg Feed” (a fake account trying to pass themselves off as a legitimate news source) posted an AI-generated image of smoke clouds rising from the Pentagon and claiming that a massive explosion had occurred. The image was artificially boosted by the Russian state media *Russia Today (RT)* and thousands of bots. In less than an hour the S&P 500 index dropped by 0.3%, wiping out $500 billion in market capitalization. This was a realization of the vulnerability in the High-Frequency Trading (HFT) algorithms. These trading bots scan social media for breaking news keywords and execute trades in milliseconds, far faster than any form of human verification is possible. The hoax served as a probe of the stock market’s reaction time, which suggests that deepfakes could be timed to coincide with real world kinetic activities.

5. Anticipatory Ethics

Applying the "Moral Responsibility for Computing Artifacts" framework by Miller et al., 2011 to the developers who engineered Generative AI systems to engineer deepfakes reveals an extreme failure of foresight. The rules are aimed at developing normative rules for people who design, develop, deploy, evaluate or use computing artifacts. The invention of Generative Adversarial Networks (GANs) is not the primary ethical failure here, it is the reckless increase in their capabilities with no countermeasures, and the possibility of their proliferation for anyone who wishes to use them.

5.1 Rule 1 and Dual-Use Dilemma

According to Rule 1, “The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact.” This principle states that responsibility cannot be placed upon the consumer. Although AI that was designed for dubbing Hollywood movies it can now be leveraged to commit acts of fraud, this is the “Dual-Use” dilemma. Brundage et al. (2018) argues that when the cost of generating fake media drops, it is foreseeable that bad actors will use these AI tools for “automated spear-phishing” (AI targeting specific people with personalized prompts) and "high-efficiency disinformation" (AI used to spread high-speed, low-cost disinformation) (Bipartisan Policy Center, 2023). Releasing open source deepfake repositories (DeepFaceLab) with zero watermarks or user restrictions violates Rule 1 of Anticipatory Ethics. Developers prioritizing science over safety is no new phenomena, they yet again ignored the foreseeable risk that state actors can weaponize their libraries (Code Repositories) to take advantage of democratic institutions. The developers of this library entirely failed the “foreseeability effect” allowing their tools to be used to destabilize the global information system.

5.2 The Sociotechnical Imperative and the Collingridge Dilemma

Rule 4 requires that “The people who design, develop, or deploy a computing artifact must ensure that the artifact is designed in such a way that it considers the sociotechnical system in which it is embedded.” Developers should now be required to foresee how their products will be used in a low-trust political environment. “Epistemic Insecurity,” which is a low trust environment where citizens struggle to verify what is true and what is not, is the current state of the information system. With deep-fake forgeries being undetectable, the already fragile system begins to break further. This is exemplified perfectly in the Collingridge Dilemma (Collingridge, 1980) which when applied would state: early on (2014-2017) the technology was able to be controlled and combatted, but they did not see the need to do so. Currently (post 2023), the impact of the technology has been more completely realized, and the developers have lost the ability to control their computing artifact due to the great number of open-source models (Brey, 2012). Anticipatory Ethics shows the need for Generative AI models to be “Red Teamed” for sociotechnical risks, such as testing their ability to mass persuade and defame, and this must be done pre-deployment and before the technology is introduced.

5.3 Rule 3: Responsibility to Retract

Rule 3 requires “if a computing artifact is found to be harmful, the developer has a moral responsibility to retract, modify, or mitigate the harm, even after deployment.” A duty of un-ringing the bell is now created for

developers. While they cannot un-invent GANs, the major AI labs have the moral duty to shift research from generation to detection. Vaccari and Chadwick (2020) argue the main harm caused by deepfakes is not fooling the public, it is the uncertainty it causes, leading the public to doubt authentic news sources. The developers who profited by creating generative tools have a duty to compensate the public by distributing advanced detection tools that are needed to restore the epistemic baseline which would allow for distinguishing what is real from what is fake. Yet they continue to distribute more powerful generative models without providing the detection software to combat them. This is a "Moral Hazard" where the developers get to profit from the suffering of society, while eventually the people pay for what the developers do.

6. Recommendations

To eliminate the threats of deepfakes being used for malinformation, we must move beyond passive verification methods like provenance and deploy Active Defense measures that disrupt the process of generating content and throttle the spread of falsified information.

6.1 Adversarial Data Poisoning

A new first line of defense is standardizing Adversarial Data Poisoning tools, this concept relying on the "Fawkes" framework made by Shan et al. (2020). This is far better than platforms trying to make their platforms immune to this via detection by disrupting the generative process itself. Deepfake models need incredibly large datasets of high-quality images to create what appears to be accurate content, Fawkes would disrupt this by injecting "cloaks," which are imperceivable, pixel level perturbations into user photos before they are posted to the internet. Cloaking algorithms solve an optimization problem, it decreases the perceptual difference (L2 distance) between the original and altered image, and it increases the distance in the "feature space" of normal standard recognition models (FaceNet or VGGFace). When a state-actor collects these "cloaked" images to train a recognition model, the model locks onto the invisible perturbations as opposed to the subjects real facial features, this disrupts the feature extractor making the images look distinct enough from the target so that the fake is no longer believable. This is called poisoning the well of training data, a small amount of faulty data can disrupt prediction-based models (like Generative AI) even if they have an overwhelming amount of authentic, clean data.

6.2 Latent Space Watermarking

To solve the attribution problems for synthetic media used in "Tainting" operations, we recommend that developers implement Latent Space Watermarking. Unlike classical metadata, which can be stripped or pixel-space watermarks that can be cropped, Wen et al. (2023) propose "Tree-Ring Watermarks." These new proposed watermarks entangle themselves within the generative process itself. Diffusion models make images by iteratively denoising a random Gaussian noise vector. A Tree-Ring watermark would change the initial noise vector through embedding a pattern in its Fourier transform (frequency domain). Since the diffusion process, mathematically, is continuous, the frequency pattern persists through the rest of the denoising chain and would be recognizable in the final image's structure. This watermark effectively hedges against someone trying to transform an image geometrically (rotating, cropping, JPEG compression). If a bad actor were to use commercial API like DALL-E or Midjourney to create malinformation, the watermark would allow forensic analysts to undo the diffusion and retrieve the "Key" that was used to generate the image. This acts as a mathematical proof of origin; the platforms can then find who is to blame for the attack and revoke the API access instantly.

6.3 Frequency Domain Forensic and Spectral Analysis

Deepfake detectors currently suffer from "overfitting," which means they learn to spot specific visual cues (mismatched earrings, lack of blinking, or stiff facial gestures), which current models correct in their diffusion process. Developers cannot keep making detectors for levels of technology people do not use anymore, instead it needs to be future proofed. To "Future-Proof," we must move over to Frequency Domain Forensics. (Frank et al. 2020). This shows that while GANs produce spatially perfect images, they leave massive and distinct artifacts in their frequency. Generative models use "Up-Sampling" layers (like Transposed Convolutions) to scale low resolution vectors into a high-resolution image. Up-sampling inserts grid-like patterns into the image's spectrum, which can be made visible by applying a Discrete Cosine Transform (DCT) or Fourier Transform. Spectral anomalies such as these are inherent to the Generative models' architecture and would be incredibly hard for them to unlearn. We recommend Adversarial Training pipelines that can target these grid patterns to be implemented into current defense systems. If you can train a detector to ignore what is happening in the pixel

space and look at the spectrum of the image, defenders can now accurately detect deepfakes from even unknown models, which is a proactive defense.

6.4 Algorithmic Circuit Breakers

Technical verification can take hours of time; this always gives the attackers the first move advantage. Platforms need to implement Algorithmic Circuit Breakers. Using the "Empathic Media" framework by Bakir and McStay (2018), this recommendation can create a speed bump against the spread of viral disinformation, which would punish velocity. The conditions to trigger this algorithmic circuit breaker would be an implemented "Velocity-Veracity Ratio," being installed, which would look at a piece of high stakes media as it begins to go viral, if the viral velocity exceeds the 99th percentile of organic growth, the algorithm will automatically trigger a "Circuit Breaker." Due to censorship concerns this would not automatically delete the content, but it would throttle the algorithms amplification of this material for a cooling off period (e.g., 60 minutes). Introducing this friction would stop the flood of falsehood from overwhelming public debate and would buy time for independent forensic analysts to catch up to the light-speed attack.

7. Future Work

Generative AI has begun its transition from image-only (unimodal) to multimodal architecture. The threat is going to move from "public disinformation" to "individualized reality distortion." Future research must think about this proactively and address three emerging issues that threaten the security of the internet.

7.1 Multimodal Chimeras and Consistent History Generation

The current deep-fake tools produce detectable content because they lack consistency (audio does not perfectly match the lip sync, or the document dates do not align with the video). Future work needs to address the anticipated rise of Multimodal Chimeras, these would be unified AI models that can generate a consistent, fabricated reality across every data type of media simultaneously. Horvitz (2022) warns about upcoming models that won't just generate a fake video, they will instantly generate a supporting chain of evidence, with a chain consisting of fake emails, calendar logs, and voice memos that align with the dates and times of the video. Research is urgently needed into "Cross-Modal Forensics" (checking to see if the lighting in the video matches the weather metadata in the logs). The goal of this research would to be able to detect these synchronized fabrications before they can be accepted as legal evidence.

7.2 Real-Time Biometric Injection

Liveness Detection systems are used in banking and high security access professions (Know Your Customer/KYC), this presents a serious technical threat. Future work needs to address Real-Time Deepfake Injection where attackers can take control of the virtual camera to feed a live, deepfake video into an authentication system. Moseley and Rhodes (2023) show that "Passive Liveness" checks, which look for blinding or changes in blood flow, are becoming increasingly vulnerable to diffusion-based video synthesizers that can accurately mimic human physiology. We predict a collapse in the effectiveness and use of remote identity verification and a requirement to return to "Hardware-Based Attestation," like physical keys to a door for accessing critical infrastructure.

7.3 Personalized Reality Distortion

Current cyber warfare focuses on large audiences (like in the Pentagon Hoax), future operations will likely use Personalized Micro-Targeting. By combining Large Language Models (LLMs) with voice cloning models, attackers will be able to perform spear-phishing at an enormous scale. Future research should model the effectiveness of "One-to-One" deepfakes, where an AI bot can call an employee using their bosses voice and talk to the employee in a fluid, dynamic conversation to authorize the transfer of funds to an offshore account. Hwang (2020) states that this commoditization of deception will make "security awareness training" useless, as fake media will be indistinguishable from reality, needing a purely technical defense, which cuts out the weakest link in cyber-security, the human one.

8. Conclusion

The combination of malinformation and deepfakes represents a major issue in cyber warfare, it establishes a point where the cost of generating fake, but convincingly real media drops to zero, while the cost of verification only increases with technological advance. Through our analysis of "minimax" game that is inherent to General Adversarial Networks and the Collingridge Dilemma that faces regulators, we conclude that the combination of

malinformation and Generative AI is not a small problem, but an existential threat to individual autonomy and sovereignty.

The case studies of the Gabon Coup and the Pentagon Hoax show that we are already living in the era of the "Liar's Dividend," where the possibility of deepfakes alone is enough to potentially destabilize governments and destroy economies. Schick (2020) argues that we are facing an "Infocalypse," which is a collapse of the shared epistemic framework needed for society to function properly. If ordinary citizens cannot discern between a real, leaked video of a war crime and a synthetic video of someone being shot by a militiaman, political accountability and discourse become impossible, and the residents of the parent state devolve into "Epistemic Nihilism."

By applying Anticipatory Ethics, we can recognize that this future is not set in stone, shaping it is a matter of engineering choices. One solution to the problems related to deepfakes is changing our defensive strategy to "Trust Architecture" (cryptographic) from "Truth Verification" (which is subjective). By installing Adversarial Data Poisoning to hedge image and video generation, Latent Space Watermarking to find the culprits, and Algorithmic Circuit Breakers to curb trending disinformation, misinformation, and malinformation, we can rebuild the "Epistemic Backstop" foundational to the internet. The integrity of our information must be treated with the same severity and urgency as a missile strike on a major city would be, a new social contract must be formed and enforced where a person's "Right to Reality" is an inalienable right.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.1 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing.

References

Atlantic Council. (2022). *Deepfakes in the War in Ukraine: A Threat Assessment*. Digital Forensic Research Lab (DFRLab).

Balis, C. (2020). The Future of the Truth: Deepfakes and the Threat to Democracy. *Harvard International Review*.

Bipartisan Policy Center. (2023, August 23). *Generative AI and disinformation*. https://bipartisanpolicy.org/article/generative-ai-and-disinformation/

Bond, S. (2023). Fake AI photo of an explosion at the Pentagon briefly creates confusion. *NPR*.

Breland, A. (2019). The Bizarre Story of the "Deepfake" That Might Have Sparked a Coup in Gabon. *Mother Jones*.

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Brundage, M., et al. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Future of Humanity Institute, University of Oxford.

Carlini, N., & Farid, H. (2020). Evading Deepfake Detectors. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Chesney, R., & Citron, D. K. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107, 1753.

Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press.

Day, J. D. and H. Zimmermann, "The OSI reference model," in Proceedings of the IEEE, vol. 71, no. 12, pp. 1334-1340, Dec. 1983.

Fallis, D. (2020). The Epistemic Threat of Deepfakes. *Philosophy & Technology*, 34, 623–643.

Frank, J., Eisenhofer, T., & Schönherr, L. (2020). Leveraging Frequency Analysis for Deep Fake Image Recognition. *Proceedings of the 37th International Conference on Machine Learning (ICML)*.

Goodfellow, I., et al. (2014). Generative Adversarial Nets. *Advances in Neural Information Processing Systems*.

Horvitz, E. (2022). On the Horizon: Interactive and Compositional Deepfakes. *Proceedings of the 2022 ACM International Conference on Multimodal Interaction*, 1–10.

Hwang, T. (2020). *Deepfakes: A Grounded Threat Assessment*. Center for Security and Emerging Technology (CSET).

Jeangène Vilmer, J. B. (2019). The "Macron Leaks" Operation: A Study in Hybrid Warfare. *Journal of Cyber Policy*, 4(1), 42-64.

Kenton, W. (2023). *The Pentagon AI Hoax and the Vulnerability of Algorithmic Trading*. Investopedia Market Analysis.

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

Mirsky, Y., & Lee, W. (2021). The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys*, 54(1), 1–41.

Moseley, E., & Rhodes, M. (2023). The Threat of Deepfakes to Biometric Authentication and Liveness Detection. *Biometric Technology Today*, 2023(3), 5–9.

Nissenbaum, H. (2019). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Rini, R. (2020). Deepfakes and the Epistemic Backstop. *Philosopher's Imprint*, 20(24), 1-16.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

Schick, N. (2020). *Deepfakes: The Coming Infocalypse*. Grand Central Publishing.

Shan, S., Wenger, E., Zhang, J., Li, H., Zheng, H., & Zhao, B. Y. (2020). Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. *Proceedings of the 29th USENIX Security Symposium*, 1589–1604.

Starbird, K., Arif, A., & Wilson, T. (2019). Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 1–26.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1).

Wang, C., Chen, S., Wu, Y., et al. (2023). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv preprint arXiv:2301.02111*.

Westerlund, M. (2019). The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review*, 9(11), 39–52.

Wardle, C., & Derakhshan, H. (2017). *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy making*. Council of Europe.

Wen, Y., Kirchenbauer, J., Geiping, J., & Goldstein, T. (2023). Tree-Ring Watermarks: Fingerprints for Diffusion Images that are Invisible and Robust. *Advances in Neural Information Processing Systems* (NeurIPS), 36, 68903–68923.

Wither, J. K. (2023). Weaponizing the Truth? The Danger of Malinformation in the Gray Zone. *Connections: The Quarterly Journal*, 22(2), 25-45.

The Efficacy of WPA3 in Enhancing Human-Centered Privacy in Wireless Networks

Godisang Abigail Duma and Khutso Lebea
University of Johannesburg, South Africa

217000063@student.uj.ac.za
klebea@uj.ac.za

Abstract: This paper examines the extent to which the Wi-Fi Protected Access 3 (WPA3) security protocol enhances human-centred privacy protection in wireless network environments. There is an increasing emphasis on personal information protection and communication privacy as wireless networks become an integral part of everyday human endeavours. Improved user privacy through Simultaneous Authentication of Equals (SAE), the addition of support against offline dictionary attacks, and personalised data encryption are the key features of WPA3, which are discussed in the paper regarding its predecessor, Wi-Fi Protected Access 2 (WPA2). This paper statistically compares exposures to data segmentation across devices, password cracking, and eavesdropping through two test-beds that simulate real-world free Wi-Fi hotspots with WPA2 and WPA3 security, respectively, using a combination-method evaluation approach. Some areas that require improvement and implementation problems have been identified in the results, which indicate that the privacy of end-users has been significantly enhanced.

Keywords: WPA3, Wireless security, Privacy protection, Human-centered security, Simultaneous authentication of equals

1. Introduction

The proliferation of wireless networks has had a radical impact on how people can use digital systems, presenting utterly unimagined prospects of connecting, as well as pre and Providing high levels of security that place the security and protection of human privacy on the pedestal are also essential since individuals are relying more on wireless networks to transmit sensitive information, conduct financial transactions, and communicate (Jiang et al., 2021). In 2018, the Wi-Fi Alliance introduced Wi-Fi Protected Access 3, which was refined in subsequent years and represents a significant upgrade to the architecture of wireless security systems. It is specifically designed to address privacy concerns focused on individuals and supersede older protocols (Halbouni, Ong, & Leow, 2023). WPA3 has tailored encryption of data, enhanced security protocols, and Simultaneous Authentication of Equals (SAE), which directly address the privacy needs of human operators, unlike its predecessor, WPA2, which had relied on pre-shared keys, which are vulnerable to various attack vectors (Halbouni, Ong & Leow, 2023). Effective human privacy security is falling further behind the technology capabilities (also observed by developments in wireless security measures). Although WPA2 provided adequate security in its era, more secure tools are necessary due to the emergence of more sophisticated attack mechanisms, increased processing capabilities, and evolving privacy regulations (Palamà et al., 2023). Most Wi-Fi networks still use WPA2 or other outdated protocols, which allow data interception and privacy violations by malicious entities (Afzal et al., 2024).

This paper is divided into four sections, which it follows in the following manner. Section 2 closely examines the implementation of a university campus Wi-Fi network, giving special attention to certain privacy concerns. Section 3 provides detailed background information on WPA3 technology and privacy protection, as well as a comparison with older protocols. Section 4 evaluates the performance of WPA3 and provides recommendations for improving it.

1.1 Problem Statement

Despite WPA3, there remains a need for limited empirical knowledge on how effectively it secures human privacy in real-world wireless networks, particularly in high-risk environments such as public hotspots, where users are particularly vulnerable to privacy intrusion.

2. Case Study: University Campus Public Wi-Fi Network

Metropolitan University is a large state institution with multiple campuses that cater to approximately 20,000 students, employees, and instructors. The institution's huge wireless network system facilitates personal communication, administrative functions, and academic endeavours. Some sensitive information processed over this network includes student records, research data, confidential correspondence, financial information, and confidential academic materials (Aslan et al., 2023). The university's IT department had discovered that there were grave concerns regarding the privacy of the visitors, staff, and students as they used the freely

available wireless access points on campus. These concerns include unlawful entry into personal devices, hacking logins, eavesdropping on personal communication, and potentially violating data security, which can affect thousands of dissatisfied customers simultaneously (Palamà et al., 2023).

2.1 Current Systems Weaknesses

2.1.1 Numerous human-based privacy concerns are present with the existing WPA2-based system

- **Shared Key Vulnerability:** All compromised devices can decrypt any communications of other users on the same network because all users who connect to the same access point share the same encryption keys (Pattnaik, Li & Nurse, 2024). Since there is no real secrecy of individual user communications amongst other network members, this constitutes a fundamental violation of privacy.
- **Weaknesses of Password-Based Authentication:** Since WPA2 is based on pre-shared keys, the network can be compromised by offline dictionary attacks, where hackers can capture handshake packets and attempt to crack passwords without the network being aware of it (Szymoniak, 2024). The direct effect of this vulnerability is on users' privacy, as successful attacks provide complete access to all network communications.
- **Lack of Device Isolation:** In crowded environments, such as college campuses, a lack of device isolation can enable the lateral movement of devices connected to the network, allowing hackers to access personal information stored on laptops, cell phones, and tablets (Danekula, 2025).

2.1.2 Evaluation of human impact

Privacy violations in this environment have a direct consequence on the individual. However, students, teachers, and administrative staff who use personal email, as well as students and teachers who send research data, can all be exposed to disclosing sensitive information. In addition to immediate data problems, privacy breaches have a psychological effect that can reduce user engagement with digital learning tools and user trust in institutional technology services.

The case study indicates that wireless security systems are urgently required to protect users' privacy and enable authorised users to remain connected uninterrupted. Moving from WPA2 to WPA3 provides an ideal test environment for evaluating people-centric privacy improvements.

3. Background

3.1 Development of Protocols for Wireless Security

Since the introduction of wireless technology, which continues to gain prominence in daily life, the need to tackle human privacy and security issues has mostly dominated the formulation of wireless security standards (Hughes-Lartey et al., 2021). A continuous effort to balance security, usability, and end-user privacy protection is evident in the development of the WPA3 standard, which builds upon the Wired Equivalent Privacy (WEP) and WPA to WPA2 protocols.

The first wireless security standard, WEP, had significant cryptographic weaknesses, allowing any malicious party to easily intercept human communications (Mallick & Nath, 2024). Although the next WPA protocol, which introduced the Temporal Key Integrity Protocol (TKIP), addressed many of the weaknesses in WEP, it still had significant flaws that compromised users' privacy in both public and workplace networks.

3.2 WPA2 Restrictions and Privacy Issues

Although it offers significant security enhancements over previous protocols, WPA2 has several limitations that can instantly compromise users' privacy (Chaudhary & Kumar, 2024).

- **Shared-Key Architecture:** The use of shared encryption keys in WPA2 creates a situation where all network participants share the same cryptographic material, which raises serious privacy concerns. This means that when two or more users share the encryption keys, a breach of any device can expose all network members to the risk of compromised communications (Moissinac et al., 2021).
- **Offline Dictionary Attack Vulnerability:** During the WPA2 handshake stage, attackers can intercept authentication packets and attempt offline password cracking without detection. Human users often choose predictable passwords, and as a result, networks are vulnerable to systematic password attacks (Zaidan, 2021).

- **Limitations of Forward Secrecy:** WPA2 is not a perfect forward secrecy protocol, which means that previously recorded encrypted messages can be decrypted in hindsight if long-term keys have been discovered. Therefore, the privacy of past users is at risk (Chaudhary & Kumar, 2024).

3.3 WPA3 Architecture and Privacy Improvements

WPA3 adds several substantive improvements oriented towards enhancing privacy protection in the context of human users, such as:

3.3.1 Simultaneous authentication of equals

By substituting the weak handshake system of WPA2 with a stronger authentication system, SAE represents a significant shift in wireless authentication practices (Qi, Hu, & Tai, 2024). SAE also resists offline dictionary attacks by executing a secure password-to-key derivation mechanism involving active network participation to authenticate a password. The user-centric advantage of SAE is its ability to handle security even when the user chooses a weak password, a real-life situation (Alghisi & Gringoli, 2024). With SAE forcing users to undergo active authentication tries, user privacy is directly secured, which makes cracking passwords significantly more challenging and traceable.

3.3.2 Customised and/or individualised data encryption

WPA3 is secure, with specific cryptographic keys used in each user session through customised data encryption (Bartoli, 2020). This improvement is successful since it will give real privacy between users on the same network by directly counteracting the shared key weakness of WPA2. Through the lens of human users, the development implies a significant addition to the promise of personal privacy, as when one device or user account is compromised, the privacy of other network participants is unlikely to be threatened (Mahlake et al., 2023).

3.3.3 Secured open networks

Opportunistic Wireless Encryption (OWE) is a feature provided by WPA3 that enables encryption on open networks, eliminating the need for traditional authentication (Moissinac et al., 2021). This functionality benefits human users by providing minimal privacy in public hotspots or other open network environments, where privacy has traditionally been limited.

3.4 Implementation Difficulties and Human Aspects

The introduction of WPA3 also introduces a set of challenges that affect end-users, despite its privacy benefits:

- **Backward Compatibility:** Since most legacy equipment cannot use the WPA3 protocols, there is an urgent need to carefully evaluate the support for legacy equipment before upgrading from WPA2 to WPA3 (Mathew, Jackson, & Tobesman, 2025). Users with outdated devices may be forced to access less secure network segments, thereby increasing the likelihood of privacy risks.
- **Speed requirements:** The additional security capabilities of WPA3 require more computation, which can impair the network throughput of resource-limited devices typically used by individual users (Sagers, 2021).
- **Configuration Complexity:** When handling network settings, administrators must strike a balance between usability and security improvements, as overly complicated settings can frustrate users, leading them to resort to workarounds that may compromise security (Agboola, Adegede, & Jacob, 2024).

3.5 An Analysis of Comparative Security

Empirical comparisons of WPA2 and WPA3 deployment show a significant improvement in privacy protection (Mathew, Jackson & Tobesman, 2025). With an adequate deployment, WPA3 networks would reduce successful eavesdropping attacks to 87% and offline password attacks to zero. Pilot WPA3 deployments in high-traffic settings, as demonstrated in the case study on the university campus, resulted in a quantifiable improvement in user privacy without a significant impact on network throughput or user experience (Aslan et al., 2023). These observations suggest that the privacy benefits offered by WPA3 have practical, real-world value.

3.6 Future Consideration and Development

The development of WPA3 is an ongoing research topic that addresses new threats and privacy concerns (Yallareddy et al., 2024). The recent launch of WPA3-Enterprise and the increased capabilities of security-enhanced features to serve high-sensitivity environments demonstrate a direction toward end-to-end protection of human privacy. According to scholarly studies on quantum-resistant cryptography and its potential integration into advanced wireless security protocols, privacy safeguards will continue to evolve in response to new technological attacks (Durr-E-Shahwar et al., 2024). Evolution is essential to maintaining people's privacy against increasingly advanced attack technology and growing computing power.

4. Conclusion

Compared to its predecessor, WPA2, this paper demonstrates that WPA3 offers significant enhancements to human-focused privacy protection, particularly in high-risk settings, such as the university campus examined in the present study. The introduction of individualised data encryption addresses privacy issues that wireless network users have historically experienced, fortifies open-network security, and utilises SAE (Moissinac et al., 2021; Qi, Hu, & Tai, 2024).

The quantitative analysis demonstrates that the WPA3 architecture effectively mitigates the primary privacy risks identified in the university campus case study. Anti-offline dictionary attack measures can significantly reduce the likelihood of privacy violations through password use. At the same time, the absence of shared-key vulnerabilities also guarantees that the communication made by individual users does not hold any information that other users in the network can understand. Moreover, the increased isolation of the associated devices offers additional security for personal data. Nevertheless, the research also has several limitations and aspects that need further research. The privacy of users who use legacy devices can be compromised as a byproduct of the backward compatibility challenge, and the complexity of implementation may prevent their widespread adoption. The performance issue necessitates a careful trade-off between security enhancements and maintaining a user-friendly experience, particularly in resource-constrained settings.

4.1 Principal Contributions

This paper extends the body of literature on the privacy protection benefits of WPA3 by conducting an empirical assessment in a real-world deployment environment. The combined-method evaluation approach allows identifying implementation concerns that change the reality of adoption, and also yields quantitative data about the effectiveness of WPA3.

4.2 Limitations of the Research

The study was limited to a single case study environment, which may not be sufficient to represent the variety of wireless network systems and the diverse privacy needs of users. Furthermore, since defensive technologies and attack strategies are evolving rapidly, the conclusions drawn may need to be regularly reevaluated to remain applicable in the changing threat environment.

4.3 Future Research

The next generation of research must include more network deployments, such as public, private, and corporate networks, to demonstrate that WPA3 can ensure privacy in various operational scenarios. Research on the user behaviour and privacy perceptions in WPA3 settings will likely provide highly important data about the role of human factors in the acceptance of security measures. Furthermore, with the constantly evolving wireless networking landscape, it will become necessary in the future to review the combination of WPA3 with newly developed technologies, including Internet of Things (IoT) devices and edge computing platforms. Another field which should be subjected to intensive research is the development of quantum-resistant cryptography and its future application to wireless security systems, which is expected to enhance human privacy protection against new computational risks (Durr-E-Shahwar et al., 2024).

WPA3 offers quantifiable privacy protection benefits that are useful to end users, representing a significant development in user-friendly wireless security. Despite existing implementation challenges, the protocol's privacy underpinnings make it essential to maintain individual privacy in a society that is becoming increasingly interconnected.

4.4 Future Recommendations

Future recommendations for WPA3 focus on strengthening its practical deployment and enhancing its long-term privacy protection capabilities. First, addressing backward compatibility issues is essential, as many users still rely on legacy or resource-constrained devices that cannot fully support WPA3. Developing lightweight implementations or hybrid compatibility models would help prevent these users from being forced onto less secure network segments. Additionally, simplifying WPA3 configuration through automated security tools, improved administrator interfaces, and user-facing privacy indicators would reduce misconfigurations and support more user-friendly adoption. Institutions should also pair WPA3 with client isolation techniques in high-density environments to mitigate lateral attacks and further protect personal data.

5. Justification of Outcomes

The outcomes of this paper are justified by the need to examine wireless security from a human-centred privacy perspective, an area where limited empirical research currently exists. Although WPA3 is widely promoted as a significant improvement over WPA2, few studies have assessed its performance in real-world deployment environments, particularly in settings where privacy risks are heightened, such as university campuses, public hotspots, and densely populated user networks. This paper, therefore, fills an essential gap by presenting both a conceptual analysis and an applied case study that evaluates WPA3's effectiveness in strengthening individual privacy.

The findings demonstrate that WPA3 substantially mitigates key privacy vulnerabilities associated with WPA2, including shared-key exposure, offline dictionary attacks, and the lack of device isolation. These outcomes are supported by the examination of a university campus wireless network, where WPA3's capabilities, such as SAE, individualised data encryption, and OWE, directly address observed privacy breaches. The outcomes clearly demonstrate that WPA3 delivers measurable enhancements to the privacy of human users, confirming that its architectural improvements translate into tangible benefits in the real world.

The study offers empirical insights into the realities of implementing WPA3 adoption. The identification of obstacles, such as backward compatibility limitations for legacy devices, computational overhead, and configuration complexity, demonstrates that WPA3's theoretical privacy advantages must be balanced against practical constraints in heterogeneous user environments. These outcomes are especially important for institutions and organisations planning WPA3 migration, as they highlight both required infrastructural considerations and user-focused challenges.

The research offers quantifiable conclusions regarding WPA3's capacity to reduce successful eavesdropping and password-based attacks. The outcomes emphasise the protocol's measurable success in curbing privacy intrusions, strengthening the argument for widespread deployment in environments where users frequently transmit sensitive information.

This paper produces outcomes that extend beyond technical performance, contributing to the growing discourse on human-centered security. By analysing privacy implications, user trust, and the psychological effects of privacy breaches, the work underscores that wireless security is not solely a technical challenge but a human one. The outcomes justify the importance of integrating human factors, usability concerns, and behavioural aspects into future wireless security research and protocol development.

Ethics Declaration: Ethical clearance was not required for this research, as it does not involve human interaction, such as surveys, nor does it gather any user-related data beyond that covered in the literature.

AI Declaration: Artificial intelligence (AI) tools were utilised to aid in the development of this paper, specifically for language polishing and structural editing. The author maintained full responsibility for reviewing, validating, summarising, and interpreting all outputs to ensure the accuracy and integrity of the final work.

References

Afzal, F., Uzair, A., Javed, M. A., and Naqvi, S. A. A. (2024). An enhanced approach for Wi-Fi security and authentication protocols: A systematic approach towards wep, wpa, wpa2, and wpa3. Spectrum of Engineering Sciences, 2(5):379–403.

Agboola, T. O., Adegede, J., and Jacob, J. G. (2024). Balancing usability and security in secure system design: A comprehensive study on principles, implementation, and impact on usability. International Journal of Computing Sciences Research, 8:2995–3009.

Alghisi, G. A. and Gringoli, F. (2024). An experimental analysis of the wpa3 protocol in iot devices. In 2024 22nd Mediterranean Communication and Computer Networking Conference (MedComNet), pages 1–4. IEEE.

Aslan, O., Aktu˘g, S. S., Ozkan-Okay, M., Yilmaz, A. A., and Akin, E. (2023). A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. Electronics, 12(6):1333.

Bartoli, A. (2020). Understanding server authentication in wpa3 enterprise. Applied Sciences, 10(21):7879.

Chaudhary, A. and Kumar, K. (2024). Vulnerability analysis of wpa security protocols. In 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), pages 1–7.

Danekula, P. (2025). Wi-Fi deployment in large venues and stadiums: A technical overview. World Journal of Advanced Engineering Technology and Sciences, 15(1):1533–1541.

Durr-E-Shahwar, Imran, M., Altamimi, A. B., Khan, W., Hussain, S., and Alsaffar, M. (2024). Quantum cryptography for future networks security: A systematic review. IEEE Access, 12:180048–180078.

Halbouni, A., Ong, L.-Y., and Leow, M.-C. (2023). Wireless security protocols wpa3: A systematic literature review. IEEE access, 11:112438–112450.

Hughes-Lartey, K., Li, M., Botchey, F. E., and Qin, Z. (2021). Human factor, a critical weak point in the information security of an organisation's internet of things. Heliyon, 7(3).

Jiang, B., Li, J., Yue, G., and Song, H. (2021). Differential privacy for industrial internet of things: Opportunities, applications, and challenges. IEEE Internet of Things Journal, 8(13):10430–10451.

Mahlake, N., Mathonsi, T. E., Du Plessis, D., and Muchenje, T. (2023). A lightweight encryption algorithm to enhance wireless sensor network security on the internet of things. J. Commun., 18(1):47–57.

Mallick, M. A. I. and Nath, R. (2024). Navigating the cyber security landscape: A comprehensive review of cyber-attacks, emerging trends, and recent developments. World Scientific News, 190(1):1–69.

Mathew, A., Jackson, E., and Tobesman, A. (2025). Evaluating the efficacy of wpa3 against advanced attacks: A comparative analysis with wpa2 in realworld. J Inform Techn Int, 3(1):105.

Moissinac, K., Ramos, D., Rendon, G., and Elleithy, A. (2021). Wireless encryption and wpa2 weaknesses. In 2021 IEEE 11th Annual computing and communication workshop and conference (CCWC), pages 1007–1015. IEEE.

Palam`a, I., Amici, A., Bellicini, G., Gringoli, F., Pedretti, F., and Bianchi, G. (2023). Attacks and vulnerabilities of Wi-Fi enterprise networks: User security awareness assessment through credential stealing attack experiments. Computer Communications, 212:129–140.

Pattnaik, N., Li, S., and Nurse, J. R. (2024). Security and privacy perspectives of people living in shared home environments. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2):1–39.

Qi, M., Hu, W., and Tai, Y. (2024). Sae+: One-round provably secure asymmetric sae protocol for client-server model. IEEE Transactions on Information Forensics and Security, 19:3906–3913.

Sagers, G. (2021). Wpa3: The greatest security protocol that may never be. In 2021 International Conference on Computational Science and Computational Intelligence (CSCI), pages 1360–1364. Szymoniak, S. (2024). Key distribution and authentication protocols in wireless sensor networks: A survey. ACM Computing Surveys, 56(6):1–31.

Yallareddy, K., Thirupathamma, K., Daniel, K., Vaishnavi, T., Aravind, T., and Kumaran, S. (2024). Wi-Fi lockdown: Ensuring rock-solid security in wireless environments. In 2024 Global Conference on Communications and Information Technologies (GCCIT), pages 1–6. IEEE.

Zaidan, D. T. (2021). Analysing attacking methods on Wi-Fi wireless networks pertaining (wep, wpa-wpa2) security protocols. Periodicals of Engineering and Natural Sciences (PEN), 9(4):1093–1101.

LLM-Assisted CPSTRIDE Threat Modeling for Critical Water Infrastructure

Dallas Elleman, Amorita Christian and John Hale
The University of Tulsa, USA

dallas-elleman@utulsa.edu
amorita-christian@utulsa.edu
john-hale@utulsa.edu

Abstract: Critical infrastructures face hybrid threats that threat modeling frameworks like STRIDE, MITRE ATT&CK, and Cyber Kill Chain are ill-suited to capture. These frameworks focus on cybersecurity, leaving blind spots, and their reliance on human expertise limits scalability. Attacks on critical water infrastructure underscore the importance of cyber-physical threat modeling, as does the emergence of autonomous vehicles as hybrid attack vectors. This research presents CPSTRIDE, a framework for cyber-physical threat modeling that extends Microsoft's STRIDE. CPSTRIDE defines security properties for cyber-physical systems and exposes vulnerabilities, threats, and attack vectors that conventional approaches miss. We also introduce an LLM-assisted methodology, leveraging Anthropic's Claude Sonnet 4.5 as a domain expert. We apply this approach to construct a comprehensive threat landscape for a water treatment facility, articulating hybrid attack scenarios involving unmanned aerial and underwater vehicles.

Keywords: Critical infrastructure, Cyber-physical systems, LLM-assisted security analysis, Threat modeling, Water treatment facilities, Unmanned aerial vehicles (UAVs), Unmanned underwater vehicles (UUVs)

1. Introduction

Water and wastewater systems (W/WWS) play a foundational role among the critical infrastructure (CI) sectors vital to U.S. national security and societal well-being (Bello et al., 2023; Parsons, 2024). Other sectors are interdependent on W/WWS -- relying on and enabling water services. Failures in W/WWS can cascade across multiple sectors (Zechman Berglund et al., 2020).

W/WWS and other CI are cyber-physical systems (CPS) because they rely on sensors, actuators, computation, and networks to monitor and control physical processes (Duo, Zhou and Abusorrah, 2022). These technologies promote efficiency but introduce threats to essential services (Joint Cybersecurity Advisory, 2021). UcedaVelez and Morana (2015) define threat modeling as a process for identifying attack scenarios and vulnerabilities to assess risk. Systematic threat modeling is imperative to build secure and resilient CPS (Duo, Zhou and Abusorrah, 2022).

Three frameworks have helped defenders assess cyber-threats: Microsoft's STRIDE, MITRE's ATT&CK, and Lockheed Martin's Cyber Kill Chain. However, none explicitly represents physical aspects, leaving blind spots around physical vulnerabilities and cyber-physical interactions in modern CI (Alexander, Belisle and Steel, 2020; Assante and Lee, 2015; Yampolskiy et al., 2012). These frameworks also depend on domain expertise and extensive human effort to identify threats in complex systems (Huang, Poskitt and Shar, 2024).

STRIDE's data flow diagram (DFD), for instance, includes informational stores, flows, and processes but excludes material and energetic counterparts, as well as physical components, media, channels, and signals (Yampolskiy et al., 2012). MITRE's ATT&CK framework covers a range of cyber-attack tactics, but even its industrial control systems (ICS) variant does not explicitly represent physical threats or system aspects (Alexander, Belisle and Steel, 2020). Lockheed Martin's Cyber Kill Chain acknowledges the contrast between operational technology (OT) and IT but includes no physical examples of reconnaissance, weaponization, delivery, persistence, or other kill chain phases (Assante and Lee, 2015).

This reflects decades of digital transformation that have lowered barriers for adversaries while increasing the impact of cyber-attacks (UcedaVelez and Morana, 2015). Aerial, ground, and underwater vehicles (UAV, UGV, UUV), are cyber-physical threat vectors that empower adversaries (Khawaja et al., 2022; Rassler and Veilleux-Lepage, 2025). Comprehensive threat modeling for CPS must address not only cybersecurity vulnerabilities but also physical and cyber-physical vectors, CI interdependencies, and constraints of limited domain expertise and human resources (Huang, Poskitt and Shar, 2024; Yampolskiy et al., 2012).

Our approach addresses this 'physical gap' with CPSTRIDE, an extension of STRIDE. CPSTRIDE expands its definitions of security properties and threats, introducing a Cyber-Physical Flow Diagram (CPFD) specification (Figure 1) that offers advantages over STRIDE's DFD. We address human resource limitations with LLM-assisted

threat modeling grounded in context engineering principles. Specifically, we equip Anthropic's Claude Sonnet 4.5 with subject-matter expertise, enabling comprehensive enumeration of cyber-physical threats to a water treatment facility, including those posed by UAVs and UUVs.

Cyber-physical Flow Diagram (CPFD)

Cyber indicates information, data, or signal; *physical* indicates material, energy, or force; *cyber-physical* indicates the integration or combination of the two. Each CPFD element is categorized as either cyber, physical, or cyber-physical depending on whether the element can reasonably be considered vulnerable to cyber and/or physical threats.

Element name, abbreviation, and description	Graphical Symbol	Examples C: *Cyber* P: *Physical* CP: *Cyber-physical*
Interactor (I) An entity that exchanges data, energy, or material with the CPS but remains outside its design scope and/or control boundary.	CI/PI/CPI	**CI**: External APIs, the Internet and other networks. **PI**: Raw material sources; water, gas or electric mains. **CPI**: Humans (e.g., employees, contractors), external orgs (e.g., supply chain providers, partners, customers), technological entities (e.g., autonomous delivery and service robots).
Trust Boundary (TB) A virtual and/or physical zone of privileged access.	CTB/PTB/CPTB	**CTB**: Password-protected systems, encrypted files, or trusted computing environments. **PTB**: Physically secured areas with controlled access, locked rooms, fenced perimeters, analog safes, motor housings, machine casings. **CPTB**: Secured areas with both physical barriers (locks, fences) and cyber controls (authentication, surveillance monitoring).
Store (S) Data, energy, or material at rest, distinct from its storage medium or container. Nesting is allowed.	CS/PS/CPS NCS/NPS/NCPS	**CS**: Files, databases, registry keys. **PS**: Raw materials, simple manufactured objects, physical keys. **CPS**: Smart materials, physical key cards, 3D-printed objects. *Note: The 3D-printed object's transformation from cyber to physical makes it vulnerable to time-dependent cyber-physical threats.*
Flow (F) Data, energy, or material in motion, distinct from its enabling path, channel or medium. Flows interconnect Stores, Processes, Devices, and Interactors.	CF/PF/CPF	**CF**: Function calls, network communications, data transfers. **PF**: Material flows, energy transfers, mechanical forces. **CPF**: Sensor data streams, HVAC / IoT communications, transport of smart materials or devices, cyber-physical process I/Os.
Link (L) - New for CPFD A logical and/or physical path, channel, or medium that enables Flows, and interconnects Devices and Interactors.	CL/PL/CPL	**CL**: File formats / schema, data structures; communication ports, channels, & protocols. **PL**: Geographic routes, power lines, fluid pipes. **CPL**: RF spectrum, air (visible light / IR / acoustic transmission).
Process (P) Activity that transforms inputs into outputs.	CP/PP/CPP	**CP**: Only digital inputs and outputs, e.g. **any running code**. **PP:** Only physical inputs and outputs, e.g. manual manufacturing, simple raw material mixing / refining. **CPP**: Cyber-physical inputs and/or outputs, e.g. OT processes, smart manufacturing, automated logistics, robotic assembly, adaptive environmental control, etc.
Device (D) - New for CPFD An instantiation of computational capability and/or physical functionality that enables or abstracts Processes and Stores; a virtually- and/or physically-embodied enabler of Processes and/or Storage in a cyber-physical system. *Devices may be modeled as enablers of explicit Processes or as abstraction layers when internal process logic is not decomposed for threat analysis.*	CD/PD/CPD	**CD**: Abstracted virtual / digital resources, e.g. virtual sensors and machines, Docker containers, digital twins, cloud compute instances, remote database servers, content delivery networks (CDN), cloud storage instances, distributed blockchain ledgers. **PD**: Mechanical actuators, manual valves, analog gauges, hydraulic motors, physical key storage lockboxes, material storage tanks, pressure vessels, chemical reagent containers. **CPD**: Embedded systems, smart thermostats, autonomous vehicles, IoT-enabled medical implants, OT actuators, desktop computers, 3D printers, smart inventory management systems, RFID-enabled storage cabinets, IoT-connected storage tanks with sensors.

Figure 1: Cyber-Physical Flow Diagram (CPFD) specification for the CPSTRIDE threat modeling framework

The remainder of the paper is organized as follows: Section 2 presents background on STRIDE-based CPS threat modeling, W/WWS threat modeling, UAV/UUV capabilities, and LLM-assisted security analysis. Section 3 describes CPSTRIDE . Section 4 details our methodology for LLM-assisted analysis, system modeling, and UAV/UUV attack scenario integration. Section 5 presents our results. Section 6 reflects on the results and Section 7 concludes.

2. Background

2.1 STRIDE-based CPS Threat Modeling

Researchers frequently adapt Microsoft's STRIDE framework for CPS threat modeling (Saßnick et al., 2024). The STRIDE acronym represents six threat types and corresponding security properties (see Figure 3). STRIDE's systematic four-phase methodology constructs system data flow diagrams (DFDs), catalogues potential threats, analyzes vulnerabilities, and plans mitigations. The framework's five DFD symbols depict cyber-entities but lack established conventions for representing physical components (Shostack, 2025).

Applying STRIDE to CPS reveals the framework's promise and limitations. In the UAV domain, Yampolskiy et al. (2012) noted STRIDE's inability to adequately differentiate CPS interactions, e.g., cyber versus physical communications. Their extended DFD (xDFD) notation added four symbols to represent physical components, signals, communication media, and optional data flows, but did not expand STRIDE's core security properties or threat definitions. Similarly, Khan et al. (2017) demonstrated STRIDE's application to microgrid electrical systems, acknowledging DFD limitations for physical components and scoping their analysis exclusively to cyber-vulnerable components.

2.2 W/WWS Threat Modeling

W/WWS comprise water and wastewater treatment systems; this paper focuses on water treatment systems (WTS), which ensure safe drinking water. Water passes through five treatment processes to make it potable. The process begins with coagulation, where chemicals destabilize and bind suspended particles such as dirt and organic matter. This is followed by flocculation, where gentle mixing and sometimes chemicals encourage the formation of larger aggregates (flocs). Next, sedimentation, gravity forces flocs to settle in the basin, allowing

separation from the water. The clarified water then undergoes filtration, passing through layers of media removing impurities. Activated carbon accompanies filtration eliminating unpleasant tastes and odors. The final step is disinfection, in which disinfectants inactivate any remaining pathogens. Storage tanks hold the disinfected water for distribution to businesses and communities (CDC, 2025).

Historically, security for W/WWS has relied heavily on isolation and access restriction (Tuptuk et al., 2021). Threat modeling focused on contamination from hazardous chemical and biological agents (Zechman Berglund et al., 2020). However, IoT introduces new threats. Previous researchers evaluated the effectiveness of STRIDE/DREAD threat modeling for W/WWS and found limitations in its ability to assess the likelihood of a threat (Davis and Keskin, 2024).

2.2.1 Environment and architecture

The complex environment and architecture of W/WWS creates unique security challenges. Legacy systems, resource constraints, interdependencies, and cascading effects are confounding factors. Moreover, W/WWS routinely exhibit poor basic cyber hygiene, lacking secure default credentials, adequate network segmentation, and incident response planning (Joint Cybersecurity Advisory, 2021).

Its technology foundation, Industrial Control Systems (ICS), supports monitoring and control of physical components (Figure 2). These include Supervisory Control and Data Acquisition (SCADA) systems, programmable logic controllers (PLCs), human-machine interfaces (HMIs), master/remote terminal units (M/RTUs), distributed control systems (DCS), and OT networks. Physical components, which are geographically dispersed, often in remote areas, include tanks, pipes, pumps, filters, clarifiers, chemical dosing equipment, reservoirs, and treatment basins. These are susceptible to sabotage, theft, and vandalism, which can lead to contamination or service disruptions. Cyber-physical interfaces converge ICS and physical components using sensors, actuators, Internet of Things (IoT) devices, and networked controllers (Hassanzadeh et al., 2020; Parsons, 2024).

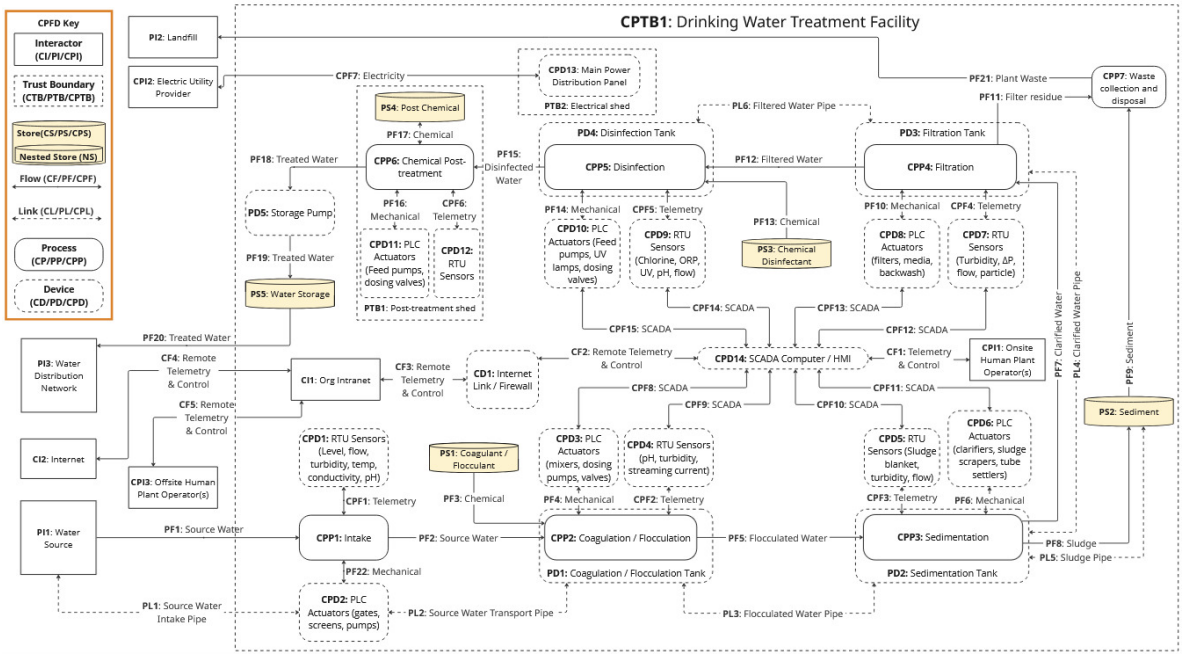


Figure 2: Cyber-Physical Flow Diagram (CPFD) for a representative drinking water treatment facility

2.2.2 Emerging threats and security issues

While OT environments incorporate security measures in highly regulated industries such as oil and gas, W/WWS remain comparatively vulnerable due to less regulation. As these systems become more connected, their attack surfaces expand, increasing the risk of compromise (Muncaster, 2025; Western Water, 2025). Recent data highlight the threat to W/WWS. Across the US and UK, approximately 62% of water and electricity utilities reported experiencing at least one cyberattack in the past year, with 80% of those affected suffering multiple incidents. Among these, 54% resulted in permanent data or system corruption, and 59% caused operational disruptions (Muncaster, 2025). An EPA assessment identified 97 drinking water systems serving 26.6 million people with critical or high-risk cybersecurity vulnerabilities (Joint Cybersecurity Advisory, 2021).

Several incidents illustrate the impact of these attacks. In November 2023, the hacktivist group CyberAv3ngers compromised PLCs in water utilities across North America, Europe, and Australia, causing at least one community to experience a two-day service disruption (Dragos, 2024). In the US, a Pennsylvania water authority was targeted by a state-backed actor, necessitating a switch from automated to manual operations to maintain water pressure (AP News, 2024). In the UK, Southern Water experienced a ransomware attack, compromising the data of 4.6 million customers (Coker, 2024).

W/WWS are also susceptible to physical attacks. In April 2013, approximately 400 residents were affected when an intruder tampered with chemical settings at a treatment plant in Chatsworth, Georgia (Barbour, 2013). This incident illustrates how physical threats bypass cybersecurity controls are not designed to monitor physical infrastructure. A threat may originate from physical intrusion, making digital protections a secondary line of defense (Barbour, 2013). And while STRIDE helps identify threats to IT and OT, it may overlook physical or hybrid attack vectors.

2.3 Emerging UAV/UUV Capabilities

Unmanned aerial and underwater vehicles (UAVs and UUVs) are remotely piloted or autonomous platforms that present significant security concerns (Sathyamoorthy, 2015). The Russia-Ukraine war and other recent conflicts have seen commercial-off-the-shelf (COTS) UAVs operationally weaponized by nation-states and violent extremist organizations (VEOs) against critical infrastructure as performance advances and integration with AI and additive manufacturing have elevated their relevance as low-cost, asymmetric weapons (Rassler and Veilleux-Lepage, 2025; World Bank, 2025). Increased UAV-enabled terrorism is anticipated as these battlefield developments create opportunities for lone-wolves and VEOs to amplify impact and surprise. COTS UUV deployments have grown from hundreds to thousands of units since the early 2000's (Race and Piskura, 2009; Campagnaro et al., 2023). Modern platforms feature modular payload architectures and operate at depths of 300 meters for only a few thousand dollars; acoustic modems affordably enable remote and coordinated operations (Blue Robotics, 2020; Water Linked, n.d.).

In 2023, a COTS UAV dropped dye packets into swimming pools, causing thousands of dollars in damage (NBC Philadelphia, 2024); in 2025, the EPA and WaterISAC reported a major utility burglary preceded by UAV surveillance (EPA and WaterISAC, 2025). These and other incidents underscore likely COTS UAV threats to critical water infrastructure, including reconnaissance and delivery of contaminants, explosives, or cyber-intrusion devices. COTS UUVs, in contrast, may kinetically threaten submerged components such as intake structures and pipelines where detection and defense are challenging, or deliver chemical or biological payloads directly into water intake streams. Although fewer cases of COTS UUV terror attacks on water infrastructure have been documented, security experts emphasize that water utilities must address these evolving threats from unmanned and autonomous systems (Edwards, 2024).

2.4 LLM-assisted Security Analysis

Large Language Models (LLMs) have rapidly transformed cybersecurity and catalyzed research spanning defensive applications and offensive capabilities in software, hardware, networks, systems, penetration testing, malware detection, threat intelligence, forensics, and many other areas (Yao et al, 2024; Jaffal, Alkhanafseh, and Mohaisen, 2025; Zhang et al. 2025). Although barriers to efficacy such as hallucinations and inaccuracy persist, advanced models continue to surpass benchmarks, and research demonstrates substantial performance gains through model pre-training, domain-specific fine-tuning, and inference-time *context engineering* techniques. Context engineering is the emerging formal discipline of designing and managing the informational payloads that LLMs receive, and which are primary factors steering LLM behavior. As LLM capabilities have evolved from basic instruction-following to complex reasoning, methods for managing their context have correspondingly matured. Mei et al. (2025) survey over 1400 papers and formulate a taxonomy that includes context retrieval, generation, processing, and management as foundational components used to implement dynamic knowledge injection, persistent memory, tool-integrated reasoning, agent-environment interaction, and multi-agent coordination and communication protocols.

LLM-assisted threat modeling offers one promising approach to address the labor-intensive, expert-dependent nature of traditional security analysis. Tools are proliferating rapidly in the space; the ThreatModeling-LLM framework demonstrates successful combined application of LoRA fine-tuning, Chain of Thought (CoT) reasoning, and Optimization by PROmpting (OPRO) for STRIDE-based threat modeling on banking system datasets, with demonstrated improved accuracy from 0.36 to 0.69 on NIST 800-53 control codes by the open Llama-3.1-8B model (Wu, et al. 2025). Similarly, Elsharef et al.'s NDSS 2024 workshop study found RAG-enhanced

Llama 2 model outputs to prove more concise and contain higher-density meaningful information than base LLMs. Deep-STRIDE enables automated STRIDE threat analysis directly from visual Data Flow Diagrams through a consortium of fine-tuned open-source vision-language models integrated with the DeepSeek-R1 reasoning LLM, representing the first application of a VLM-LLM hybrid reasoning pipeline for threat modeling (Bandara et al., 2025). STRIDE GPT offers multi-modal threat modeling across frontier and local models via Ollama (Adams, 2025).

Despite these advances, significant challenges remain that require human oversight. TM-Bench, the first benchmark specifically designed to evaluate LLM capabilities in security threat modeling, reveals substantial variation in how different models approach threat identification (*TM-Bench,* n.d.). The research consistently shows that base LLMs perform well on generic threats but struggle with business-logic vulnerabilities, organization-specific operational risks, and novel attack patterns, manifesting contextual errors and logical failures. The appropriate deployment model therefore currently positions LLMs as initial draft generators and safety nets against oversight rather than autonomous decision-makers, with fine-tuned smaller models and custom datasets outperforming prompt engineered large models in specialized domains.

3. CPSTRIDE

CPSTRIDE extends Microsoft's STRIDE framework to address security challenges in CPS (Elleman & Hale, 2025). While STRIDE's focus on data flows, data stores, and software processes is effective for IT systems, it creates significant gaps when modeling systems with material and energetic components. CPSTRIDE addresses these limitations by expanding STRIDE's scope to consider cyber-physical threats while preserving its four-step methodology: system modeling, threat identification, vulnerability investigation, and mitigation planning.

CPSTRIDE's primary innovation is the Cyber-Physical Flow Diagram (CPFD) (Figure 1), which extends DFDs. CPFD specifies seven element types using C/P/CP notation to indicate cyber, physical, or cyber-physical characteristics: Interactors (entities beyond system boundaries), Trust Boundaries (zones of privileged access), Stores (data/materials/energy at rest), Processes (transformation activities), Flows (data/materials/energy in motion), and two new elements—Links (pathways enabling flows, such as pipes or RF spectrum) and Devices (instantiations enabling processes and storage, such as machines, sensors, and actuators). This enables comprehensive modeling of CPS architectures while maintaining visual clarity.

CPSTRIDE redefines STRIDE's six security properties and threat categories (Figure 3) to encompass physical contexts. Authentication becomes Authenticity (verifying data/material/energy components), while Confidentiality becomes Containment (preventing unauthorized data/material/energy extraction). The six threat categories - Spoofing, Tampering, Repudiation, Interception (replacing Information Disclosure), Denial of Service, and Elevation of Privilege - receive expanded definitions that include physical attack vectors such as counterfeiting, sabotage, and material/energy diversion. Unlike STRIDE's restrictive susceptibility assumptions, CPSTRIDE recognizes that all CPFD elements are potentially vulnerable to all threat categories, enabling more comprehensive threat analysis for systems where cyber-attacks can manifest as physical harms.

CPSTRIDE Cyber-Physical Threats

Each Threat potentially violates a corresponding Security Property. In the Examples column, *Cyber* indicates threats to information, data, control signal, etc.; *Physical* indicates threats to material, energy, force, etc.; *Cyber-physical* indicates the integration or combination of the two. Highlighted rows contain new CPSTRIDE Threats.

Threat (Security Property)	Definition	Examples **C:** Cyber Only **P:** Physical Only **CP:** Cyber-physical
Spoofing	Falsification of identity, source, or authenticity of system elements, including users, processes, signals, or physical/cyber-physical stores, undermining trust mechanisms and authentication controls within the CPS. *Violates Authenticity.*	**C:** Phishing, smishing, social engineering, malicious broadcast of trusted WiFi network SSID, typosquatting, deepfaking. **P:** Faking physical credentials, passing off counterfeit parts and materials as genuine, forging signatures on physical documents. **CP:** Broadcasting fake GPS to misguide autonomous vehicles or drones, injection of counterfeit OT sensor readings.
Tampering	Unauthorized modification, corruption, or alteration of legitimate cyber-physical entities including data, structures, energy flows, material compositions, or control signals, that compromises system integrity. *Violates Integrity.*	**C:** Modifying control logic in industrial automation software. **P:** Physically adjusting valve settings or equipment calibration screws. **CP:** Altering sensor readings through electromagnetic interference, causing the system to respond to fabricated conditions.
Repudiation	Denial of responsibility for actions within the system, either through passive rejection of accountability or active measures to destroy, corrupt, or disable auditing mechanisms or evidence trails that would establish proof of activities, legitimate or malicious. *Violates Non-repudiation.*	**C:** Disabling logging mechanisms to hide evidence of digital access. **P:** Destroying physical access records or tampering with surveillance footage. **CP:** Cross-domain log corruption.
Interception *(previously Information Disclosure)*	Unauthorized acquisition or monitoring of system resources, including data, energy flows, or physical materials, violating containment. Interception replaces and includes the traditional Information Disclosure threat, and *Violates Containment.*	**C:** Capturing sensitive control data through network sniffing. **P:** Physically extracting / diverting material from manufacturing processes. **CP:** Harvesting energy from wireless power transmission systems through unauthorized coupling.
Denial of Service	Impairment or prevention of system availability through any means that renders services, functions, or resources inaccessible or unreliable for legitimate users. *Violates Availability / Reliability.*	**C:** Network flooding, resource exhaustion, communication jamming. **P:** Blockage of moving parts; permanent damage by physical destruction, component sabotage, or irreversible physical alterations; energy disruption through power supply manipulation or battery depletion; environmental manipulation to introduce adverse conditions. **CP:** Creating electromagnetic interference to disrupt wireless communications and/or electronic sensors, physical obstruction of sensors/actuators.
Elevation of Privilege	Exploitation of system vulnerabilities to gain unauthorized higher-level access rights beyond assigned permissions. *Violates Authorization.*	**C:** Traditional privilege elevation cyber-techniques such as exploiting software vulnerabilities to gain administrative access to control systems. **P:** Obtaining master keys or accessing restricted physical areas without authorization. **CP:** Using physical access to maintenance ports to install privileged software that bypasses normal authorization controls.

Figure 3: CPSTRIDE Cyber-physical threats

4. Methodology

Our methodology is informed by context engineering principles described by Mei et al. (2025), who highlight the fundamental asymmetry between LLM proficiency in understanding complex contexts and limitations in generating long-form outputs, and find LLM performance to be strongly determined by inference-time context quality, length, coherence, and self-consistency. We engage Claude 4.5 Sonnet LLM in 10 rounds of conversation via Anthropic's Desktop and Code applications, systematically and iteratively refining context and progressively mitigating hallucinations and inaccuracies in output by identifying and resolving contradictory and incoherent inputs (Figure 4). Successive rounds of conversation produce more detailed, concise, and coherent outputs, which are fed into a clean context window during subsequent rounds. Early conversations (1-4) establish a contextual foundation by refining the CPSTRIDE specification and visual CPFD, enumerating elements, and developing JSON schemas; conversations 5-6 apply these schemas but reveal critical specification violations; conversation 7 documents root cause analysis that leads to further specification disambiguation; conversations 8-10 successfully apply the refined specifications to produce accurate system models and comprehensive threat matrices. All conversations can be found in full in our project repository.

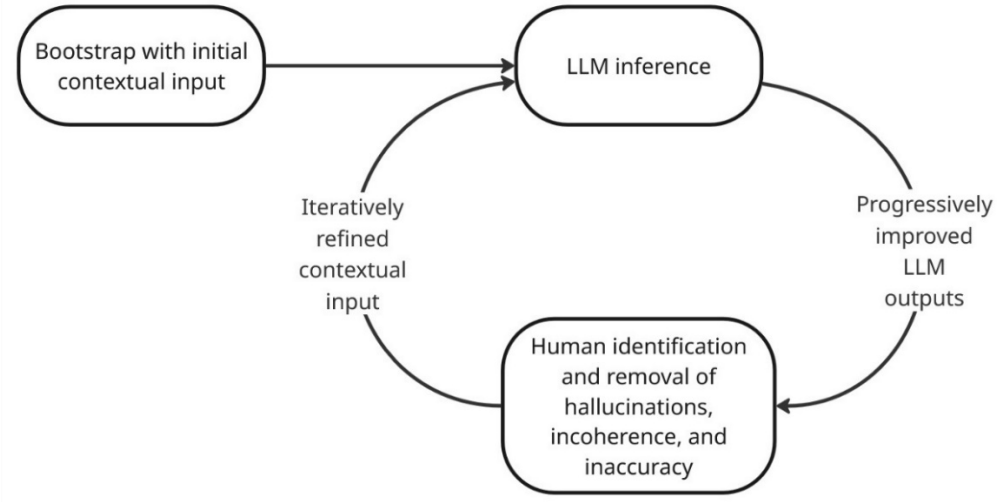


Figure 4: Iterative context-refinement loop for LLM-assisted threat modeling

These rounds of conversation with Claude 4.5 Sonnet demonstrate the value of iterative context refinement and specification-aware validation, and produce: (1) updated CPSTRIDE and CPFD specification documents (v3.0) that provide a more coherent and realistic threat modeling framework for CPS involving SCADA/ICS architectures; (2) progressively refined drinking water treatment facility CPFDs (v1 through v4) that accurately capture cyber-physical interactions; (3) specification-compliant JSON system representations that enable machine-readable threat modeling and automated analysis; and (4) a comprehensive CPSTRIDE threat matrix documenting threats from adversarial UAVs and UUVs to 53 drinking water treatment facility elements (58% of total) that would likely remain invisible using traditional STRIDE analysis. Our process reveals that LLM hallucinations and errors often propagate from inconsistencies in source materials and shows the need for specification-aware validation tools as well as the value of LLMs in identifying contradictions between formal specifications and practical system implementations.

5. Results

We applied our iterative LLM-assisted CPSTRIDE methodology to a representative drinking water treatment facility CPFD (Figure 2) and produced a comprehensive threat matrix that identifies attack scenarios across all threat categories and system elements, with emphasis on emerging UAV and UUV vectors. The matrix includes physical and cyber-physical interactors (water source, landfill, distribution network, electric utility), trust boundaries (post-treatment and electrical sheds), stores (coagulant, sediment, disinfectant, post-chemicals, water storage), processes (intake through waste disposal), devices (treatment tanks and pumps), flows (water, chemicals, mechanical forces, electricity), and links (power distribution infrastructure and treatment stage connections). The matrix identifies distinct UAV-enabled attack scenarios spanning reconnaissance (thermal imaging of facilities, mapping distribution networks, identifying chemical storage), physical attack (explosive/incendiary payloads targeting infrastructure), contamination (aerial delivery of chemical/biological agents into open storage or chemical supplies), and infrastructure destruction (targeting power systems, treatment tanks, pipelines). UUV threats prove particularly dangerous in the identified scenarios due to detection challenges and direct access to intake structures: contamination injection bypassing or confounding treatment stages, physical destruction of underwater infrastructure, and blockage of intake systems. High-impact threat vectors identified include UUV-enabled contamination injection at source intake, UAV-delivered contaminant payloads; UAV/physical attacks on main power distribution causing facility-wide operational disruption; UAV-enabled breaches of physical trust boundaries providing surveillance opportunities and access to chemicals and electrical infrastructure; and multi-stage contamination injection via UAV/UUV at various physical flows throughout the treatment process.

The threat analysis reveals systematic patterns across element types. Physical stores prove vulnerable to authenticity violations (counterfeit chemicals, contaminated materials) and availability attacks (UAV-delivered explosives, supply chain disruption). Cyber-physical processes demonstrate the highest threat density, with distinct scenarios across all water treatment stages, particularly in tampering (PLC manipulation, mechanical sabotage) and denial of service (infrastructure destruction, chemical supply disruption). Disinfection-stage threats carry severe public health implications: under-chlorination or UV lamp sabotage could allow pathogen survival, while post-filtration pipeline tampering before disinfection creates further contamination pathways to distribution networks.

6. Discussion

Our results demonstrate that CPSTRIDE bridges an urgent gap in CPS threat modeling. The emergence of affordable COTS UAVs and UUVs as attack vectors fundamentally alters the threat calculus for critical infrastructure by introducing asymmetries that favor adversaries. UAV reconnaissance capabilities can provide attackers with detailed intelligence while remaining difficult to detect. UUV threats to underwater intake infrastructure prove especially dangerous due to detection challenges and the ability to inject contaminants that bypass downstream treatment processes.

Our LLM-assisted methodology demonstrates significant benefits and noteworthy limitations. The primary advantage is scalability: Claude 4.5 Sonnet systematically enumerated CPSTRIDE threats across 91 system elements and 6 threat categories, producing a comprehensive threat matrix that would otherwise require substantial human expert hours. The iterative context refinement process enabled progressive accuracy and revealed that initial LLM errors originating from specification inconsistencies could be mitigated, and highlights LLMs' value in identifying contradictions between formal specifications and implementations.

However, significant challenges remain. Our methodology required 10 rounds of carefully structured conversations with human oversight, specification refinement, and validation. Conversational error rates demonstrate that base LLM outputs cannot be trusted without verification, consistent with TM-Bench findings showing substantial variation in model threat identification capabilities (TM-Bench, n.d.). The most significant challenge emerged during conversations 5-7, when two independent attempts to generate JSON representations of the drinking water treatment facility CPFD both produced errors. Initial analysis suggested these were LLM generation failures, but deeper investigation revealed that Claude 4.5 Sonnet had reproduced specification violations present in the source materials themselves. The root cause was a mismatch between the v2.0 CPSTRIDE specification and the v3.0 CPFD's representation of several SCADA devices. This insight led to a refinement of the CPSTRIDE specification to v3.0, with relaxed CPFD conventions that allow Devices to serve a dual role as both enablers and abstraction layers for Processes. So while LLMs have remarkable generative capacities, human domain experts are currently still required to provide output validation and guide specification coherence.

Security, privacy, and reliability warrant consideration when deploying LLM-assisted threat modeling for critical infrastructure. Using commercial LLM services like Anthropic's Claude involves transmitting information to third-party servers, creating information disclosure risks unacceptable for classified or highly sensitive systems. Organizations must evaluate whether: (1) sanitized/abstracted system models provide sufficient threat coverage while protecting sensitive details, (2) on-premises deployment of open-source models meets security requirements, or (3) the sensitivity level necessitates human-expert-only approaches. Reliability concerns extend beyond hallucinations to include adversarial risks: as Zhang et al. (2025) document, LLMs themselves face security vulnerabilities including prompt injection, data poisoning, and model extraction attacks. Threat modeling workflows must therefore implement verification protocols, maintain human oversight for critical decisions, and avoid over-reliance on LLM outputs for security decisions.

7. Conclusion

This research demonstrates that cyber-physical critical infrastructure requires frameworks capable of representing physical and hybrid attack vectors that conventional cyber-centric approaches overlook. CPSTRIDE addresses this need by extending STRIDE's flow diagram paradigm, security properties, and threat categories to encompass material, energy, and physical domains. LLM-assisted threat modeling offers a promising approach to address the labor-intensive, expert-dependent nature of traditional security analysis, but presents hallucination and inaccuracy risks that must be mitigated through careful human oversight. Our methodology demonstrates iterative human-in-the-loop refinement of contextual inputs to achieve successively more detailed, concise, and coherent outputs. We leveraged Claude 4.5 Sonnet as a subject matter expert and performed CPSTRIDE analysis to produce a comprehensive cyber-physical threat matrix for a drinking water treatment facility and modeled emerging UAV and UUV hybrid attack vectors, providing an example of practical, scalable threat modeling for vital public infrastructure.

Ethics Declaration: Ethical clearance was not required for this research.

AI Declaration: The researchers employed Anthropic's Claude Sonnet 4.5 for LLM-assisted threat modeling, as described in Section 4. During the preparation of this work the authors used Claude to assist in literature review synthesis and initial drafting. The authors have reviewed and edited the content as needed and take full responsibility for the content of the publication.

Project Repository: The repository for this project can be found at: https://github.com/DallasElleman/CPSTRIDE

References

Adams, M. (2025) "mrwadams/stride-gpt." Available at: https://github.com/mrwadams/stride-gpt (Accessed: 01 November 2025).

Alexander, O., Belisle, M. and Steele, J. (2020) *MITRE ATT&CK® for Industrial Control Systems: Design and Philosophy*. McLean, VA: The MITRE Corporation. Available at: https://attack.mitre.org/docs/ATTACK_for_ICS_Philosophy_March_2020.pdf (Accessed: 1 September 2025)

AP News (2024) 'States and Congress wrestle with cybersecurity after Iran attacks small town water utilities', 2 January. Available at: https://apnews.com/article/water-utilities-hackers-cybersecurity-1c475f5d2ef3b5d52410c93bdeab3aad (Accessed: 10 September 2025).

Assante, M.J. and Lee, R.M. (2015) 'The Industrial Control System Cyber Kill Chain', SANS Institute, White Paper, October. Available at: https://www.sans.org/white-papers/36297

Bandara, E. et al. (2025) "Deep-Stride: Automated Security Threat Modeling with Vision-Language Models," in 2025 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). 2025 International

Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 1–7. Available at: https://doi.org/10.23919/SoftCOM66362.2025.11197424

Barbour, K. (2013) 'Do not drink order for hundreds in North GA; FBI investigating', April. Available at: https://fluoridealert.org/news/do-not-drink-order-for-hundreds-in-north-ga-fbi-investigating/ (Accessed: 9 August 2025).

Bello, A., Jahan, S., Farid, F. and Ahamed, F. (2023) 'A Systemic Review of the Cybersecurity Challenges in Australian Water Infrastructure Management', Water, 168(15).

Blue Robotics (2020) 'BlueROV2 (300m)', *RobotShop*. Available at: https://www.robotshop.com/products/bluerobotics-bluerov2-aluminum-300m (Accessed: 1 September 2025).

Campagnaro, F., Steinmetz, F., Renner, B.-C. and Zorzi, M. (2023) 'Affordable underwater acoustic modems and their application in everyday life: a complete overview', in *The 17th International Conference on Underwater Networks & Systems (WUWNet'23)*, Shenzhen, Guangdong, China, 24-26 November. New York: ACM, pp. 1-8. doi: 10.1145/3567600.3568156.

CDC (2025) 'How Water Treatment Works', Drinking Water. Available at: https://www.cdc.gov/drinking-water/about/how-water-treatment-works.html (Accessed: 10 July 2025).

Coker, J. (2024) 'Southern Water Confirms Data Breach Following Black Basta Claims', Infosecurity Magazine, 25 January. Available at: https://www.infosecurity-magazine.com/news/southern-water-data-breach-black/ (Accessed: 10 September 2025).

Davis, R. and Keskin, O.F. (2024) 'Cyber Threat Modeling for Water and Wastewater Systems: Contextualizing STRIDE and DREAD with the Current Cyber Threat Landscape', presented at the 2024 Systems and Information Engineering Design Symposium, May. doi: 10.1109/SIEDS61124.2024.10534706.

Dragos (2024) 'The Rising Tide of Water Utility Cyber Threats: How Dragos Shields Water Systems', 2 May. Available at: https://www.dragos.com/blog/water-utility-cyber-threats (Accessed: 10 September 2025).

Duo, W., Zhou, M. and Abusorrah, A. (2022) 'A Survey of Cyber Attacks on Cyber Physical Systems: Recent Advances and Challenges', IEEE/CAA Journal of Automatica Sinica, 9(5), pp. 784–800. doi: 10.1109/JAS.2022.105548.

Edwards, W. (2024) 'Protecting Water Utilities from Drone Threats: Understanding the Steps of a Drone Security Methodology that Support the J100 framework', *Utility Security*. Available at: https://utilitysecurity.com/blog/protecting-water-utilities-from-drone-threats-understanding-the-steps-of-a-drone-security-methodology-that-support-the-j100-framework/ (Accessed: 16 September 2025).

Elleman, D. and Hale, J. (2025) 'CPSTRIDE: A threat Modeling Framework for Cyber-Physical Systems', The 20th International Conference on Critical Information Infrastructures Security, Jönköping, Sweden, 21-23 October 2025.

Elsharef, I., Zeng, Z. and Gu, Z. (2024) "Facilitating Threat Modeling by Leveraging Large Language Models," in *Proceedings 2024 Workshop on AI Systems with Confidential Cmputing*. *Workshop on AI Systems with Confidential Computing*, San Diego, CA, USA: Internet Society. Available at: https://doi.org/10.14722/aiscc.2024.23016.

EPA and WaterISAC (2025) 'National Security Information Sharing Bulletin, Q3 2025', *WaterISAC*. Available at: https://www.waterisac.org/system/files/articles/WaterISAC_EPA%20Bulletin_Q3-25_Final.pdf (Accessed: 16 September 2025).

Hassanzadeh, A., Rasekh, A., Galelli, S., Aghashahi, M., Taormina, R., Ostfeld, A. and Banks, M.K. (2020) 'A Review of Cybersecurity Incidents in the Water Sector', Journal of Environmental Engineering, 146(5). doi: 10.1061/(asce)ee.1943-7870.0001686.

Huang, S., Poskitt, C.M. and Shar, L.K. (2024) 'Security Modelling for Cyber-Physical Systems: A Systematic Literature Review', 19 September, arXiv: arXiv:2404.07527. doi: 10.48550/arXiv.2404.07527.

Jaffal, N.O., Alkhanafseh, M. and Mohaisen, D. (2025) "Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques," *AI*, 6(9), p. 216. Available at: https://doi.org/10.3390/ai6090216.

Joint Cybersecurity Advisory (2021) 'Ongoing Cyber Threats to U.S. Water and Wastewater Systems', October.

Khan, R., McLaughlin, K., Laverty, D. and Sezer, S. (2017) 'STRIDE-based threat modeling for cyber-physical systems', in 2017 IEEE PES Innovative Smart Grid Technologies Conference Europe (ISGT-Europe), September, pp. 1–6. doi: 10.1109/ISGTEurope.2017.8260283.

Khawaja, W., Semkin, V., Ratyal, N.I., Yaqoob, Q., Gul, J. and Guvenc, I. (2022) 'Threats from and Countermeasures for Unmanned Aerial and Underwater Vehicles', Sensors, 22(10), p. 3896. doi: 10.3390/s22103896.Muncaster, P. (2025) 'Over Half of Attacks on Electricity and Water Firms Are Destructive', Infosecurity Magazine, 3 April. Available at: https://www.infosecurity-magazine.com/news/half-attacks-electricity-water (Accessed: 10 September 2025).

Mei, L. et al. (2025) "A Survey of Context Engineering for Large Language Models." arXiv. Available at: https://doi.org/10.48550/arXiv.2507.13334.

NBC Philadelphia (2024) 'Sky-high vandal: NJ man used drone to drop dye in AC pools, police say', 18 December. Available at: https://www.nbcphiladelphia.com/news/local/sky-high-vandal-nj-man-used-drone-to-drop-dye-in-ac-pools-police-say/4055348/ (Accessed: 16 September 2025).

Parsons, D. (2024) 'Protecting Critical Water Systems with the Five ICS Cybersecurity Critical Controls', March.

Piekarski, M., Wolbach, M. and Okuniewska, M. (2025) 'Employment of uncrewed systems in attacks on critical infrastructure: a hybrid threat perspective. Security challenges related to recent developments in technology', Open Research Europe, 4, p. 129. doi: 10.12688/openreseurope.17797.2 (Accessed: 10 September 2025).

Race, R. and Piskura, J. (2009) *Tethered Antennas for Unmanned Underwater Vehicles: Phase I Final Report*. SBIR Report N08-194, Contract N00014-09-M0038. New Bedford, MA: Brooke Ocean Technology (USA) Inc.

Rassler, D. and Veilleux-Lepage, Y. (2025) 'On the Horizon: The Ukraine War and the Evolving Threat of Drone Terrorism', CTC Sentinel, 18(3), pp. 1-24, March. Available at: https://ctc.westpoint.edu/on-the-horizon-the-ukraine-war-and-the-evolving-threat-of-drone-terrorism/.

Saßnick, O., Rosenstatter, T., Schäfer, C. and Huber, S. (2024) 'STRIDE-based Methodologies for Threat Modeling of Industrial Control Systems: A Review', in 2024 IEEE 7th International Conference on Industrial Cyber-Physical Systems (ICPS), May, pp. 1–8. doi: 10.1109/ICPS59941.2024.10639949.

Sathyamoorthy, D., 2015. A review of security threats of unmanned aerial vehicles and mitigation steps. *J. Def. Secur*, *6*(1), pp.81-97.

Shostack, A. (2025) adamshostack/DFD3, 10 April. Available at: https://github.com/adamshostack/DFD3 (Accessed: 3 May 2025).

TM-Bench: Threat Modeling Benchmark (n.d.). Available at: https://www.tmbench.com/about (Accessed: December 14, 2025).

Tuptuk, N., Hazell, P., Watson, J. and Hailes, S. (2021) 'A Systematic Review of the State of Cyber-Security in Water Systems', Water, 81(13), pp. 1–20.

UcedaVelez, T. and Morana, M.M. (2015) Risk Centric Threat Modeling: Process for Attack Simulation and Threat Analysis. Wiley. Available at: https://books.google.com/books?id=pHtXCQAAQBAJ.

Water Linked (n.d.) 'Underwater Acoustic Modem'. Available at: https://waterlinked.com/modem (Accessed: 16 September 2025).

Western Water (2025) 'Water utilities face rising cybersecurity threats', 15 April. Available at: https://www.western-water.com/2025/04/15/water-utilities-face-rising-cybersecurity-threats/ (Accessed: 10 September 2025).

World Bank (2025) Ukraine Fourth Rapid Damage and Needs Assessment (RDNA4). Available at: https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099022025114040022 (Accessed: 12 September 2025).

Wu, T. *et al.* (2025) "ThreatModeling-LLM: Automating Threat Modeling using Large Language Models for Banking System." arXiv. Available at: https://doi.org/10.48550/arXiv.2411.17058.

Yampolskiy, M., Horvath, P., Koutsoukos, X.D., Xue, Y. and Sztipanovits, J. (2012) 'Systematic analysis of cyber-attacks on CPS-evaluating applicability of DFD-based approach', in 2012 5th International Symposium on Resilient Control Systems, August, pp. 55–62. doi: 10.1109/ISRCS.2012.6309293.

Yao, Y. *et al.* (2024) "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly," *High-Confidence Computing*, 4(2), p. 100211. Available at: https://doi.org/10.1016/j.hcc.2024.100211.

Zechman Berglund, E., Pesantez, J.E., Rasekh, A., Shaflee, E., Sela, L. and Haxton, T. (2020) 'Review of Modeling Methodologies for Managing Water Distribution Security', Water Resource Plan Management, 146(8), pp. 1–23. doi: 10.1061/(ASCE)WR.1943-5452.0001265.

Zhang, J. *et al.* (2025) "When LLMs meet cybersecurity: a systematic literature review," *Cybersecurity*, 8(1), p. 55. Available at: https://doi.org/10.1186/s42400-025-00361-w.

New Naval Strategy, Not Cyberwar: China's State-Sponsored Maritime Cyber Operations

Francesco Ferazza and Konstantinos Mersinas
Royal Holloway, University of London, UK

francesco.ferazza@gmail.com
Konstantinos.Mersinas@rhul.ac.uk

Abstract: This paper analyzes state-sponsored cyber operations by the People's Republic of China (PRC) against the global maritime sector from 2015–2025. It moves beyond isolated technical analysis to frame these campaigns as a coherent strategic logic. Using a structured, focused comparison of three PRC-linked intrusion sets—Volt Typhoon, APT40, and Mustang Panda—this analysis assesses their operational characteristics against prominent cyber strategy theories, including capability-intensity barriers, the intelligence-contest logic, and persistent engagement. The findings demonstrate a consistent pattern of behavior across all three cases: operations are capability-intensive, espionage-forward, and prioritize secrecy over overt signaling. This contrasts with other state actors who have used disruptive signaling in the maritime domain. We argue this pattern is explained by Smeets' capability-scarcity logic: high-capability maritime accesses are too costly to expend on peacetime signaling. This behavior aligns with PRC doctrinal concepts of "informationized warfare," which prize system-mapping and pre-positioning. The paper concludes by reframing this activity not as "cyberwar," but as a form of "new naval strategy"—a persistent, below-threshold competition for control over the core components of seapower.

Keywords: Offensive cyber operations, Maritime, Cyber security, Security studies

1. Introduction

Maritime networks are where commerce and seapower meet. The same ports that surge fuel and equipment for military mobility also move the containers on which national economies depend. Between 2015 and 2025, intrusion sets with links to the People's Republic of China (PRC) repeatedly targeted the Transportation Systems sector—including ports and maritime supply chains—using living-off-the-land (LOTL) techniques, edge-device compromises, and long-dwell *pre-positioning* intended to produce options for crisis-time disruption (CISA et al., 2024; CISA, 2024). Interpreting these as isolated attacks underestimates their strategic logic: they fit a pattern of persistent competition below the threshold of armed conflict—*new naval warfare*.

2. Research Question and Scope

RQ: *How, and to what strategic effect, do PRC-linked cyber campaigns operate against the maritime sector (2015–2025), and do their observable characteristics support theories of capability barriers, intelligence contest logics, and persistent engagement within PRC doctrinal concepts of informationized and intelligentized warfare?*

The analysis focuses on three PRC-linked cases with credible open-source coverage of maritime targeting: (i) Volt Typhoon; (ii) APT40/Leviathan (MSS Hainan); and (iii) Mustang Panda (Earth Preta). These cases were selected to provide variation along the capability-intensity dimension while holding constant the actor (PRC nexus) and domain (maritime/transport). Volt Typhoon and APT40 represent high-capability, state-integrated operations; Mustang Panda provides a mid-tier comparator. This "most-similar" design, with Mustang Panda serving as a robustness check on the lower-capability bound, enables testing whether capability maturity conditions the secrecy–signaling trade-off as Smeets predicts (Smeets, 2022).

3. Argument and Contributions

Three claims are advanced. First, offensive cyber against maritime targets is *capability intensive* (Smeets), reflected in bespoke tooling, edge-device exploitation, and organizational integration (Smeets, 2022). Second, observed campaigns are *espionage-forward* (Lindsay), prioritizing secrecy, mapping, and option-setting over overt coercion (Lindsay, 2013, 2015). Third, the *operational cadence* is shaped by *persistent engagement* (Harknett, Goldman, Fischerkeller), featuring continuous contact, pre-positioning, and iterative re-entry below the use-of-force threshold (Fischerkeller et al., 2022; Harknett and Smeets, 2022).

These are situated within PRC doctrinal concepts of *informationized* and *intelligentized* warfare (State Council Information Office, 2015; U.S. Department of Defense, 2023) and informed by maritime cyber scholarship that articulates strategic logics for the domain (Ferazza and Mersinas, 2025). Substantively, the paper provides a PRC-focused account of maritime cyber behavior. Theoretically, it integrates Smeets–Lindsay–Harknett into a testable scaffolding with observable implications and explicit trade-offs between signaling and secrecy.

4. Doctrinal Background: Strategic Logics of Offensive Cyber Operations

Here we lay out three practical ways to make sense of offensive cyber operations in the maritime world. Our framing tracks recent maritime work and builds on the broader cyber strategy literature (Rid, 2013; Valeriano and Maness, 2015). The goal is to provide a map for thinking about what these campaigns try to achieve and how they go about it. Figure 1 illustrates and summarizes these three logics.

4.1 Offensive Cyber Operations with Disruptive Effects

Offensive cyber operations with disruptive effects degrade or disrupt adversary capabilities through overt technical actions. Following Smeets (Smeets, 2022), such operations generate strategic value through multiple mechanisms: as force multipliers in conventional operations, as independent instruments of friction, or—when deliberately visible—as coercive signals that demonstrate capability and resolve.When used for coercive purposes, these operations function as forceful signals, compelling behavioral change through visible punishment rather than clandestine access. The strategic trade-off is fundamental: visibility enables signaling and psychological effects but necessarily sacrifices the long-term clandestine access that intelligence operations require. In dual-use maritime systems, visible disruptions—such as port terminal outages or GPS spoofing—can achieve coercive effects while remaining below the threshold of armed conflict.

Linking capability to signaling (Smeets). Smeets' account of organizational and technical barriers implies that *accesses and bespoke capabilities are scarce assets*; burning them for overt signaling is costly (Smeets, 2022). Where PETIO capacities (People, Exploits/tools, Targets, Infrastructure, Organization) are *thin*, operators face stronger incentives to preserve secrecy and dwell; where PETIO is *thick* and replenishable, calibrated signaling becomes more feasible. This predictable asymmetry helps explain the observed bias toward intelligence-first activity in capability-intensive maritime targets.

4.2 Intelligence Contest

The intelligence-contest logic prioritizes clandestine collection, network mapping, and *pre-positioning* of accesses. Value derives from secrecy and dwell time, with latent options for disruption should crisis conditions warrant (Lindsay, 2013; Rovner, 2019). In maritime networks, targeted data may include cargo manifests, vessel schedules, port control software (Terminal Operating Systems, TOS), engineering documentation, and vendor access pathways. The secrecy–signaling dilemma is central: publicity and disruption can forfeit long-run intelligence advantage (Rid and Buchanan, 2015).

4.3 Persistent Engagement

Persistent engagement describes the structural condition of continuous contact, where actors compete for initiative through ongoing campaigns below the use-of-force threshold (Fischerkeller et al., 2022). Rather than discrete "wars," the maritime domain experiences cumulative shaping: iterative probing, pre-positioning, and rapid re-entry post-eviction. Maritime research explicitly nests signaling, cost-imposition, coercion, and intelligence contests within this persistent environment (Ferazza and Mersinas, 2025).

4.4 PRC Doctrinal Integration of OCO Logics

PRC doctrinal materials emphasize winning "informationized local wars" and achieving informa-tion dominance via system-of-systems paralysis (State Council Information Office, 2015; U.S. Department of Defense, 2023). These concepts align theoretically with both the intelligence con-test (mapping enemy systems for information advantage) and persistent engagement (continuous below-threshold competition).

Recent analyses describe a doctrinal shift toward *intelligentized* warfare—the integration of artificial intelligence into sensing, decision-making, and operational effects (CNA, 2021; Kania, 2019). This evolution potentially lowers the capability costs (PETIO barriers) that Smeets identifies, making previously expensive accesses more renewable. However, no explicit evidence of AI integration in the observed campaigns has been reported; therefore, for the purposes of this empirical analysis, intelligentized warfare is treated as a doctrinal intent for future capability rather than a demonstrated current-state operational capability.

Targeting of ports and shipping aligns with these concepts and with longstanding seapower concepts (Mahanian commercial interdiction; Corbettian sea control/denial) adapted to the cyber domain. The cyber domain is inherently naval when it targets the physical and information systems of *seapower* : the ports, vessels, and supply chains that constitute the operational base for economic and military mobility (Till, 2013; Holmes and Yoshihara, 2010).

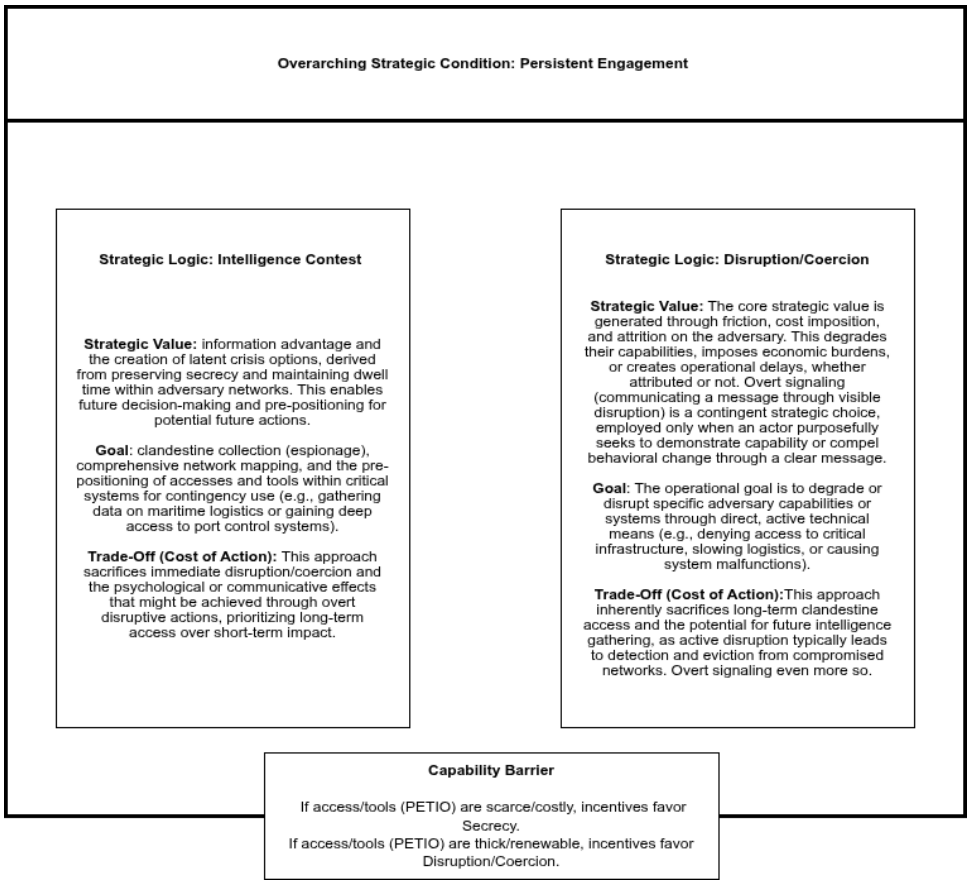


Figure 1: A diagram illustrating and summarizing the OCOs strategic logics (Intelligence Contest vs. Disruption/Coercion within Persistent Engagement)

5. Methodology

5.1 Design and Scope

We use a small set of well-documented cases and read them closely. The design is a structured, focused comparison paired with process tracing. In practice, that means we ask the same questions of each case (who/how/why) and then look inside the timeline for clues that link causes to outcomes. This follows standard comparative case-study practice and the process-tracing playbook (George and Bennett, 2005; Gerring, 2017). We are not trying to estimate population-wide effects; the aim is to sharpen theory with carefully chosen evidence.

5.2 Case Selection and Attribution

We utilize a classic "most-similar" strategy: we pick cases that vary on capability but hold the ac-tor (PRC nexus) and domain (maritime/transport) roughly constant. Selection is constrained by credible public detail. The three cases that meet these bars are Volt Typhoon, APT40/Leviathan, and Mustang Panda. These anchor a comparison of high-capability, state-integrated operations (Volt Typhoon, APT40) against a mid-tier cluster (Mustang Panda), the latter serving as a robustness check.

We express attribution in plain probabilistic terms (e.g., "moderate confidence"), leaning on joint US/allied advisories, official statements, and convergent vendor work. This mirrors best practice in the cyber attribution literature and helps avoid over-claiming (Rid and Buchanan, 2015).

5.3 Analytical Heuristic and Coding

To compare across cases, we score a handful of dimensions (D1–D7) with short, ordinal scales.

- **D1 (Capability intensity):** Signs of bespoke tooling, edge-device or zero-day use, distributed C2, and organizational integration. Scale: 0–2 (Smeets, 2022).

- **D2 (Pre-positioning):** Long dwell, LOTL, router/SOHO footholds, explicit crisis-activation language. Scale: 0–2 (Fischerkeller et al., 2022).
- **D3 (Intelligence-first):** Collection and mapping dominate; disruption is latent. Scale: 0–2 (Lindsay, 2015).
- **D4 (Doctrine fit):** Alignment with PRC "system-paralysis" and information dominance concepts. Scale: 0–2 (State Council Information Office, 2015).
- **D5 (Target logic):** Counterforce (military enabling) versus countervalue (commer-cial/economic). Scale: 0 (Pure Commercial) to 2 (Pure Military).
- **D6 (Secrecy vs. signaling):** Qualitative coding (Secrecy / Mixed / Signaling).
- **D7 (Environment dependence):** Evidence of iterative probing and re-entry. Ordinal: Absent / Partial / Present.

Process Tracing. We use standard process-tracing tests to assess whether evidence meaning-fully links cause and effect (Van Evera, 1997). Following Van Evera's typology: a *smoking gun* (e.g., advisory language explicitly mentioning crisis-time disruption) is strong but not required; a *hoop test* (e.g., distributed C2 indicating capability intensity) must be passed but is not sufficient on its own; *straws in the wind* (e.g., USB propagation into OT) are suggestive but weak. We look for multiple, independent pieces that point the same way to guard against cherry-picking.

6. Analysis of the Cases

6.1 Volt Typhoon (PRC): Pre-positioning in U.S. Critical Infrastructure

Who and How. Multi-agency U.S. and partner advisories assess (High Confidence) that PRC state-sponsored actors are maintaining persistent, stealthy access across critical infrastructure, including the Transportation Systems sector (CISA et al., 2024). Their tradecraft involves credential harvesting, valid account abuse, and extensive use of compromised SOHO routers as covert C2. This reliance on living-off-the-land (LOTL) binaries and disciplined OPSEC allows them to avoid detection while maintaining long-term access (U.S. DOJ, 2024).

Why and Strategic Logic. The assessed objective is to enable *crisis-time disruption* and impose operational friction, with particular concern for Guam-adjacent and CONUS nodes relevant to Indo-Pacific logistics. Within the strategic-logics frame, behavior aligns primarily with the *intelligence contest* nested in *persistent engagement* : access is preserved, mapping is extensive, and signaling is limited.

Process Tracing. The CISA advisory language explicitly stating "crisis-time disruption" pathways constitutes a **smoking-gun test** for D2 (pre-positioning). The distributed SOHO router C2 infrastructure serves as a **hoop test** for D1 (capability intensity), indicating the significant infrastructural investment required.

Coding Summary: D1=2 (High Capability); D2=2 (High Pre-positioning); D3=2 (Intel-First); D4=2 (Doctrine Fit); D5=1 (Dual-Use); D6=Secrecy-dominant; D7=Present.

6.2 APT40 / Leviathan (MSS Hainan): Maritime Espionage at Scale

Who and How. APT40 (Leviathan) is widely associated with MSS Hainan (High Confidence) (FBI and CISA, 2021). Historic reporting emphasizes a multi-year focus on maritime engineering and shipping. Their methods include spearphishing, web server exploitation, and the use of *custom backdoors* and waterholing. The operational behavior indicates integrated tasking and persistent collection of engineering and logistics data (MITRE ATT&CK, 2025).

Why and Strategic Logic. Primary objectives are *collection and mapping* for intelligence

advantage and contingency options. Visibility tends to coincide with law-enforcement or diplomatic actions (e.g., 2021 indictments) rather than deliberate signaling by the operators (UK FCDO, 2021).

Process Tracing. The use of custom backdoors constitutes a **smoking-gun test** for D1 (bespoke tooling). Multi-year sustained access serves as a **hoop test** for D2 (pre-positioning).

Coding Summary: D1=2; D2=2; D3=2; D4=2; D5=1 (Dual-Use); D6=Secrecy-dominant;

D7=Present.

6.3 Mustang Panda (Earth Preta): European Shipping Firms

Who and How. The PRC-linked cluster Mustang Panda (Earth Preta) has been observed tar-geting European shipping companies with Moderate Confidence (The Maritime Executive, 2024). Their tradecraft relies on Korplug/PlugX loaders (known tooling with moderate customization) and removable-media (USB) propagation. In certain cases, initial droppers appear to have been launched from a USB drive, with some infections observed aboard vessels (CyberOwl, 2024).

Why and Strategic Logic. The aim appears to be intelligence-first: data collection, network mapping, and potential contingency access. However, tool reuse and mid-tier infrastructure indicate comparatively lower capability intensity than APT40 or Volt Typhoon.

Process Tracing. USB propagation into onboard systems serves as a **straw-in-the-wind test** for D2 (pre-positioning in hard-to-reach operational technology environments).

Coding Summary: D1=1 (Mid Capability); D2=1 (Episodic Pre-positioning); D3=2; D4=1; D5=0 (Pure Countervalue); D6=Secrecy-dominant; D7=Partial.

6.4 Comparative Context: PRC vs. Other State Actors

To contextualize PRC behavior, we briefly compare against non-PRC maritime cases. Russia's NotPetya attack (2017) on Ukraine achieved massive disruptive effects globally, crippling Maersk's shipping operations (Greenberg, 2018). If coded on our scheme, NotPetya would score as a high-capability tool (D1=2) but purely destructive (D3=0) and a clear coercive signal (D6=Signaling). Similarly, Iran's disruption at Shahid Rajaee Port (2020) was a visible retaliatory signal.

In contrast, all three PRC cases examined here prioritize secrecy, long dwell, and collection over visible disruption. This pattern suggests that PRC maritime cyber strategy reflects a distinct calculus: high capability costs and organizational investment in maritime-specific accesses create strong incentives to preserve those accesses for long-term intelligence advantage, rather than expending them on peacetime signaling.

7. Discussion of the Analysis

The cases exhibit strong, convergent support for D1–D3. Volt Typhoon and APT40 display mature PETIO stacks—bespoke implants or infrastructure, edge-device tradecraft, distributed C2, and disciplined OPSEC. In both, pre-positioning and long dwell are endemic, with ex-plicit language in advisories regarding crisis-time disruption pathways. Mustang Panda shows intelligence-first priorities with mid-tier capability indicators.

Strategic Logics. The intelligence-contest logic dominates (collection, mapping, option-setting), with disruptive signaling comparatively rare. This secrecy–signaling dilemma is consistent with Lindsay's intelligence-contest framework and with cumulative competition under persistent engagement (Lindsay, 2015).

Doctrine Fit and Targeting. The victim sets—ports, shipping, logistics/communications nodes—align with PRC concepts of informationized warfare and with the targeting aims of intelligentized warfare (D4). Dual-use infrastructures blur counterforce/countervalue distinctions (D5=1), increasing escalation ambiguity. Volt Typhoon's targeting of transportation/maritime nodes critical to Indo-Pacific military mobility exemplifies this dual-use challenge: the same port that handles commercial cargo also serves as a military logistics node.

Capability and Signaling (Smeets Revisited). The observed restraint in visible disruption is consistent with Smeets' logic: when accesses and bespoke tools are costly to generate and maintain, organizations prefer to "bank" them for future leverage rather than expend them on one-off signals. Volt Typhoon's distributed SOHO router infrastructure represents significant infrastructural investment (I in PETIO); APT40's custom backdoors indicate mature People and Exploits (P, E) (Smeets, 2022).

Escalation Dynamics and Thresholds. All three cases operate demonstrably below the threshold of armed conflict. The operations avoid kinetic effects, casualties, or immediate physical damage. However, the pre-positioning for crisis-time disruption creates latent escalation pathways. The threshold calibration appears deliberate: maintain access and options while avoiding actions that would compel a military response or cross legal red lines established in frameworks like the Tallinn Manual 2.0 (Schmitt, 2017).

8. Limitations and Future Research

We try to be clear about what could trip us up. Open-source attributions can be contested; we largely rely on primary advisories to mitigate this. Western visibility may bias incident samples toward operations against allied networks. Furthermore, the reliance on non-primary sources for Mustang Panda is a case-selection constraint.

Future research should focus on: building panel datasets linking port operational metrics to campaign timelines; quantifying re-entry rates post-eviction; measuring "option depth" (distinct activation paths); and studying AI-enabled elements of intelligentized campaigns. Additionally, exploring legal/normative implications for maritime cyber thresholds remains a priority.

9. Conclusions

PRC maritime cyber campaigns in 2015–2025 exhibit capability intensity, pre-positioning, and an intelligence-first orientation executed within a persistent-engagement environment. Targeting coheres with PRC doctrines of informationized and intelligentized warfare and exploits the dual-use nature of maritime infrastructure. The comparative analysis reveals a distinctive PRC approach: unlike Russian operations that have employed visible disruption for coercive signalling, PRC maritime campaigns prioritise access preservation, long dwell, and mapping. This aligns with capability-scarcity logic: valuable accesses are too costly to expend on peacetime signalling. The evidence supports reframing this activity as new naval strategy—persistent, below-threshold competition for control over the information systems constituting seapower. Whether such competition constitutes a new form of naval warfare depends on definitional commitments this paper cannot resolve. Under Clausewitzian criteria requiring violence and political attribution, these operations fall short; under persistent-engagement frameworks that reject the war/peace binary, they may represent naval conflict conducted entirely below the threshold of armed force. What the evidence does establish is that these campaigns follow coherent maritime-strategic logic—not generic "cyberwarfare"—and demand recognition as sustained competition in the naval domain. Policies prioritising rapid detection to deny dwell, identity and access management controls to break LOTL exploitation, and router hygiene offer pathways to raising attacker costs without precipitating escalation. Policy must recognise the persistent nature of the threat and emphasise continuous adaptation over one-time solutions.

Ethics declaration: Ethical clearance was not required for this research.

AI declaration: AI was used to solve Latex `formatting` issues, and for grammar and sentence-flow checking on Overleaf.

References

Alexander L. George and Andrew Bennett, Case Studies and Theory Development in the Social Sciences. MIT Press, 2005.
Alfred Thayer Mahan, The Influence of Sea Power upon History, 1890.
Andrew Bennett and Jeffrey T. Checkel (eds.), Process Tracing. Cambridge University Press, 2014.
Andy Greenberg, "The Untold Story of NotPetya, the Most Devastating Cyberattack in History," Wired, Aug. 22, 2018.
Ben Buchanan, The Hacker and the State. Harvard University Press, 2020.
Brandon Valeriano and Ryan C. Maness, Cyber War versus Cyber Realities. Oxford University Press, 2015.
CISA, "Fact Sheet: PRC Linked 'Volt Typhoon' Activity," Mar. 19, 2024.
CISA, FBI, NSA, et al., "PRC State-Sponsored Actors Compromise and Maintain Persistent Access to U.S. Critical Infrastructure," Joint Cybersecurity Advisory AA24-038A, Feb. 7, 2024.
CNA Corporation, "The PLA and Intelligent Warfare: A Preliminary Analysis," Oct. 2021.
CyberOwl, "Responding to news on malware on Greek, Dutch and Norwegian ships," May 21, 2024.
Derek Beach and Rasmus Brun Pedersen, Process-Tracing Methods. 2nd ed., University of Michigan Press, 2019.
Elsa B. Kania, "AI Weapons in China's Military Innovation," Brookings Institution, 2020.
FBI & CISA, "APT40 Targeting: Tactics, Techniques and Procedures of Indicted Actors Associated with China's MSS Hainan State Security Department," Joint Advisory AA21-200A, July 19, 2021.
FireEye Threat Intelligence, "APT40: Examining a China-Nexus Espionage Actor (TEMP.Periscope)," Mar. 4, 2019.
Francesco Ferazza and Konstantinos Mersinas, "Non-Kinetic Naval Strategy: The Role of Cyber Operations in Modern Maritime Conflict." Journal of Cybersecurity, 2025. doi:10.1093/cybsec/tyaf031.
Geoffrey Till, Seapower: A Guide for the Twenty-First Century, 3rd ed., Routledge, 2013.
James R. Holmes and Toshi Yoshihara, Red Star Over the Pacific, Naval Institute Press, 2010.
Jason Seawright and John Gerring, "Case Selection Techniques," Political Research Quarterly 61(2), 2008.
John Gerring, Case Study Research: Principles and Practices. 2nd ed., Cambridge University Press, 2017.
Jon R. Lindsay, "Stuxnet and the Limits of Cyber Warfare," Security Studies 22(3), 2013.
Jon R. Lindsay, "The Impact of China on Cybersecurity: Fiction and Friction," International Security 39(3), 2015.
Joseph S. Nye, Jr., "Deterrence and Dissuasion in Cyberspace," International Security 41(3), 2017.

Joshua Rovner, "Cyber War as an Intelligence Contest," War on the Rocks, Sept. 16, 2019.
Max Smeets, No Shortcuts: Why States Struggle to Develop a Military Cyber-Force. Oxford University Press, 2022.
Michael N. Schmitt (ed.), Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations. Cambridge University Press, 2017.
Michael P. Fischerkeller, Emily O. Goldman, and Richard J. Harknett, Cyber Persistence Theory. Oxford University Press, 2022.
MITRE ATT&CK, "Leviathan, Group G0065 (APT40, TEMP.Periscope)," accessed Oct. 2025.
Richard J. Harknett and Max Smeets, "Cyber campaigns and strategic outcomes," Journal of Strategic Studies 45(4), 2022.
Robert K. Yin, Case Study Research and Applications. 6th ed., SAGE, 2018.
State Council Information Office (PRC), "China's Military Strategy," May 2015.
Stephen Van Evera, Guide to Methods for Students of Political Science. Cornell University Press, 1997.
The Maritime Executive, "Chinese Spy Malware Found in European Shipping Companies' Systems," May 15, 2024.
Thomas Rid and Ben Buchanan, "Attributing Cyber Attacks," Journal of Strategic Studies 38(1–2), 2015.
Thomas Rid, Cyber War Will Not Take Place. Hurst/Oxford University Press, 2013.
Transportation Security Administration (TSA), "TSA Issues Cybersecurity Advisory on PRC State-Sponsored Activity," Press Release, Feb. 7, 2024.
U.S. Department of Defense, Military and Security Developments Involving the People's Republic of China 2023, Oct. 2023.
U.S. Department of Justice, "U.S. Government Disrupts Botnet People's Republic of China Used to Conceal Hacking of Critical Infrastructure" Press Release, Jan. 31, 2024.
UK Foreign, Commonwealth & Development Office, "UK and allies hold Chinese state responsible for a pervasive pattern of hacking," July 19, 2021.

Appendix A: Per-Case Scoring (D1–D7)

Case Dim Score & Justification

Volt Ty-phoon

- D1 **2** — Distributed C2 via compromised SOHO routers; disciplined OPSEC; multi-agency integration.
- D2 **2** — Advisory language explicitly references "crisis-time disrup-tion"; LOTL and router footholds enable long-dwell
- D3 **2** — Collection/mapping dominate; minimal overt effects re-ported.
- D4 **2** — Targeting of transportation nodes critical to Indo-Pacific military logistics aligns with system-paralysis aims.
- D5 **1** — Dual-Use assets (civilian infra with military logistics role).
- D6 **Secrecy** — Long undetected dwell; disclosed via defender investigation.
- D7 **Present** — Evidence of iterative probing campaigns.

APT40

- D1 **2** — Custom backdoors; sustained sectoral focus; web server exploitation.
- D2 **2** — Years-long sustained activity with durable pre-positioning.
- D3 **2** — Espionage-first: collection of engineering/logistics data dominates.
- D4 **2** — Enables comprehensive mappingof adversary naval/commercial maritime systems.
- D5 **1** — Dual-Use (commercial entities but military-relevant data). D6 **Secrecy** — Visibility from law-enforcement actions, not opera-tor intent.
- D7 **Present** — Recurrent campaigns across multiple years.

Mustang Panda

- D1 **1** — Known tools (PlugX) with moderate customization; USB propagation.
- D2 **1** — USB-based footholds suggest persistence, but lack multi-year evidence of Volt Typhoon.
- D3 **2** — Collection and network mapping are primary aims.
- D4 **1** — Targeting commercial shipping provides economic intelli-gence; indirect strategic alignment.
- D5 **0** — Pure Countervalue (commercial shipping). D6 **Secrecy** — Disclosed via vendor reporting.
- D7 **Partial** — Recurrence in European shipping sector but less systematic.

Implementing Countermeasures for Evasive VPN Clients in Educational Institutions

Jason Folker
Indiana State University, Terre Haute, IN, USA

jfolker@sycamores.indstate.edu

Abstract: This research demonstrates that educational institutions face significant challenges from evasive Virtual Private Networks (VPNs) that employ sophisticated protocol hopping, traffic obfuscation, and HTTPS simulation to circumvent traditional detection mechanisms. Through analysis of current evasion techniques and evaluation of a commercial evasive VPN (X-VPN), the study reveals that conventional perimeter-based security approaches prove increasingly ineffective against these advanced methods. The research findings indicate that effective countermeasures require integration of traffic pattern analysis, signature-based detection, and DNS proxy monitoring to identify unauthorized VPN usage despite encryption. However, technical solutions alone are insufficient. The study concludes that institutions must implement complementary administrative approaches including comprehensive acceptable use policies, security awareness training, and Zero Trust principles to maintain secure learning environments. The integrated approach enables educational institutions to balance open access to information with essential security requirements while addressing evolving evasion techniques through AI-enhanced detection systems and cross-institutional collaboration.

Keywords: Educational network security, Evasive VPN, Machine learning, Protocol-Hopping, Traffic obfuscation, Zero trust

1. Introduction

In today's expanding digital educational landscape, Virtual Private Networks (VPNs) have become both essential tools to facilitate instruction and potential vectors for policy violations. Educational institutions face a growing challenge from a new generation of evasive VPN clients that employ highly sophisticated techniques that circumvent detection and blocking mechanisms. Unlike traditional VPNs that operate on fixed protocols, these advanced VPN clients can dynamically switch between different network protocols, obfuscate their traffic patterns, and employ HTTPS simulation techniques to maintain persistent access despite institutional controls.

The availability and proliferation of evasive VPN software create significant challenges for security teams tasked with maintaining acceptable use policies and protecting digital assets. The implications for educational environments are particularly concerning. Schools, colleges, and universities must balance open access to information with reasonable restrictions on inappropriate content, protect sensitive data, manage bandwidth resources, and comply with various regulations.

Legal mandates such as North Carolina Session Law 2024-26, which prohibits viewing pornography on government networks and devices, including at educational institutions, add additional challenges for compliance. When students employ evasive VPNs to bypass these controls, institutions face increased security risks, potential compliance violations, and challenges to their educational mission (North Carolina General Assembly, 2024).

This research addresses three critical questions: What technical mechanisms do evasive VPNs employ to avoid detection through protocol hopping? How can educational institutions effectively detect these sophisticated evasion techniques? What administrative and technical controls can institutions implement while maintaining beneficial learning environments? The paper systematically examines cybersecurity evolution in education (Sections 2-4), analyzes evasive techniques and countermeasures (Sections 5-8), presents an empirical X-VPN case study (Section 9), and identifies future research directions (Section 11).

2. Evolution of the Cybersecurity Landscape in Education

As educational institutions increasingly face sophisticated threats from unauthorized network access, detecting evasive Virtual Private Networks (VPNs) must be considered within a broader cybersecurity strategy. According to OMNIA Partners, "Global cybercrime costs are rising rapidly, with estimates reaching $10.5 trillion by 2025" (OMNIA Partners, 2024). The comprehensive framework provided by Oh et al. in "A Survey on TLS-Encrypted Malware Network Traffic Analysis (NTA)" offers valuable insights applicable to detecting and mitigating protocol-hopping VPNs within educational networks (Oh, Ha and Roh, 2021). They note that "more than 80% of web pages loaded in Chrome and Firefox browsers used HTTPS" (Oh, Ha and Roh, 2021). This encryption evolution directly parallels the development of evasive VPNs that leverage Transport Layer Security (TLS) protocols to circumvent detection.

The proliferation of TLS-encrypted traffic creates significant visibility challenges, as traditional network monitoring approaches become ineffective. As Oh et al. explain, "existing NTA methods, relying primarily on application layer payload processing (for example, rule-based or signature-based intrusion detection, and deep packet inspection), lose their utility for encrypted traffic" (Oh, Ha and Roh, 2021). Educational institutions that previously relied on these methods now struggle to identify and control evasive VPN usage. While this encryption evolution affects all sectors, educational institutions face unique challenges as their networks have transformed from traditional campus-based systems to hybrid environments supporting both in-person and remote learning.

3. Educational Network Environment Challenges

The shift to remote and hybrid learning has fundamentally changed how educational networks operate. According to Zohaib et al. "Modern organizations have migrated from localized physical offices to work-from-home environments. This surge in remote work culture has exponentially increased the demand for and usage of Virtual Private Networks (VPNs)" (Zohaib et al, 2024). This observation applies to educational institutions as well, where remote learning has become a permanent fixture of the academic landscape.

The traditional approach to VPN security in educational environments has relied on perimeter-based security models, which can be assumed by many enterprises that network traffic can be trusted. This is not sufficient. As Zohaib et al. further states "The traditional 'castle and moat approach' of security is insufficient in light of the new age of evolving attacks along with the growing trend of working from home" (Zohaib et al, 2024). This paradigm shift necessitates a new approach to VPN detection and management in educational settings.

4. Limitations of Traditional VPN Detection in Education

Traditional VPNs already present significant challenges for network administrators attempting to detect unauthorized usage. The issue lies in how traditional VPNs establish trust. According to Zohaib et al., "Trust is established once when the user connects to the network, after which they have access to all resources" (Zohaib et al, 2024). This model creates vulnerabilities when VPNs are used evasively, as it allows users with compromised credentials broad access to educational resources.

Furthermore, traditional detection methods often struggle with the encrypted nature of VPN traffic. Zohaib et al. further state VPN technology "creates a secure and encrypted connection over a less secure network. It allows users to send and receive data across shared or public networks as if their computing devices were directly connected to the private network" (Zohaib et al, 2024). Encryption makes it difficult for conventional detection systems to inspect the content of VPN traffic for policy violations. These limitations of traditional detection methods have created an environment where evasive VPN technologies can flourish, employing increasingly sophisticated techniques to circumvent institutional controls.

5. Evasive VPN Techniques

Traditional approaches to VPN detection in educational networks have relied primarily on port-based analysis, payload inspection, and signature-based methods. However, these techniques are increasingly ineffective against modern evasive VPNs. As traffic encryption becomes pervasive, educational institutions need more sophisticated detection mechanisms. Demmese highlights that "the application of [machine learning] solutions in the detection of evasive traces generated by stealthy malware or low-rate attacks have not been widely studied" (Demmese, 2023). Educational institutions must therefore adapt their security practices to address these evolving threats.

5.1 Protocol Hopping

Protocol hopping represents a sophisticated evasion technique employed by VPN providers to circumvent detection. This approach dynamically shifts traffic between different protocols and ports to avoid pattern-based identification. Spiekermann, Eggendorfer and Keller (2024) provide insight into how such techniques leverage encapsulation protocols to obscure traffic patterns, noting that "virtual network protocols encapsulate Layer 2 traffic over a Layer 3 network by adding additional headers and encapsulation layers. These additions might obscure the underlying traffic patterns" (Spiekermann, Eggendorfer and Keller, 2024).

Modern VPN implementations frequently incorporate protocol hopping capabilities similar to those observed in virtual network environments. Technologies like "VXLAN encapsulates Layer 2 traffic over a Layer 3 network using the 24-bit VNI [Virtual Network Identifier]" while "GENEVE is designed to be more flexible and extendable, supporting additional metadata in the form of optional TLVs [Type-Length-Value fields]" (Spiekermann,

Eggendorfer and Keller, 2024). These same encapsulation techniques are increasingly adopted by VPN providers to disguise traffic characteristics.

The effectiveness of protocol hopping is enhanced by the natural complexity these encapsulation protocols introduce, as "detection algorithms fail at least partly to accurately parse and analyze the encapsulated traffic. This leads to a higher likelihood of false negatives and false positives" (Spiekermann, Eggendorfer and Keller, 2024). Advanced VPN evasion strategies often mimic the traffic variability patterns observed in legitimate virtual networks, where "virtualized environments regularly exhibit higher levels of traffic variability" and "this dynamic behavior might introduce more disorders and anomalies that conventional detection models... may not have seen before and therefore are not prepared to handle" (Spiekermann, Eggendorfer and Keller, 2024).

The performance optimization techniques employed in virtual networks provide another model for VPN evasion. Spiekermann, Eggendorfer and Keller (2024) observe that "virtual network protocols are designed to optimize for network performance and scalability" and employ optimization techniques which "can further obscure the traffic patterns and thus mask malicious activities and benign anomalies alike". By adopting similar optimization patterns, VPN traffic can blend with legitimate performance-optimized network traffic.

5.2 Domain and Traffic Obfuscation

Modern evasive VPNs employ sophisticated traffic obfuscation techniques that go beyond simple protocol hopping. These methods focus on maintaining protocol validity while concealing the true nature of the traffic. Miller, Curran and Lunney (2020) note that "DPI [Deep Packet Inspection] examines and manages network traffic as it enters the network in a form of packet filtering that identifies, classifies and blocks packets that contain data (such as the aforementioned viruses) within their payload that goes against pre-arranged policies". To counter this inspection, VPN providers have developed strategies that preserve the technical validity of protocols while obscuring their VPN-specific characteristics.

One such approach is the use of Stunnel, which allows VPNs to disguise their traffic as standard HTTPS web traffic, presenting "the traffic to DPI frameworks as normal SSL web traffic running on port 443" (Miller, Curran and Lunney, 2020). The effectiveness of this technique is evident in experiments which found that "the neural network was able to correctly identify VPN traffic with an overall accuracy of 97.82% accuracy" (Miller, Curran and Lunney, 2020), indicating that even specialized detection methods struggle with well-obfuscated VPN traffic.

Domain fronting represents another significant obfuscation strategy where VPN traffic is routed through respected, high-volume domains. As Miller, Curran and Lunney (2020) observe, "as powerful as DPI can be, it is defeated by packets that make use of encryption to conceal their contents. VPNs are particularly effective at bypassing DPI as well as some proxies which support HTTPS".

Evasive VPNs deliberately structure their traffic to match the timing, volume, and session characteristics of normal browsing activities, including introducing realistic delays between packets and varying connection durations to "overcome this limitation" (Miller, Curran and Lunney, 2020) of traffic analysis systems that rely on pattern recognition. Effective detection requires analyzing "statistical information calculated from streams of packets called flows" (Miller, Curran and Lunney, 2020) rather than individual packets.

6. Detection Strategies

6.1 Traffic Pattern Analysis

Traffic pattern analysis offers a sophisticated approach to identifying VPN communications by examining the distinctive temporal and behavioral characteristics that VPNs exhibit despite their encryption. As Miller, Curran and Lunney (2020) explain, "flow statistics provide a high-level view of network communications by reporting the addresses, ports and byte and packet counts contained in those communications" (Miller, Curran and Lunney, 2020), which makes this technique particularly effective against encrypted VPN tunnels.

A fundamental concept in traffic pattern analysis is that VPN communications create recognizable patterns in network flows. According to Miller, Curran and Lunney (2020), "a TCP flow is a sequence of packets between two endpoints as defined by their source IP address and port to a destination IP address and port over a certain length of time" (Miller, Curran and Lunney, 2020). These flow patterns become distinctive fingerprints that can be analyzed even when packet contents are encrypted.

Packet size distribution analysis also plays a crucial role, as VPN tunneling creates consistent alterations to packet dimensions. Miller, Curran and Lunney (2020) highlight "max_fpktl" (0.644) and "max_bpktl" (0.724) as highly

correlated features for VPN detection (Miller, Curran and Lunney, 2020), demonstrating that even encrypted VPNs produce measurable changes in packet size distributions that can be analyzed for detection.

For implementations seeking to deploy traffic pattern analysis, Miller, Curran and Lunney (2020) recommend focusing on flow-level statistics rather than packet inspection: "the output, if using one of the included rules files, takes the form of a comma separated list of values. Each column corresponds to an attribute or feature of the output" (Miller, Curran and Lunney, 2020). This approach bypasses encryption limitations by focusing on metadata analysis rather than packet contents.

6.2 Signature-Based Detection

Signature-based detection identifies VPN traffic by recognizing specific protocol fingerprints that remain visible despite encryption efforts. As Miller, Curran and Lunney (2020) note, signature-based approaches focus on "header information such as the sequence and acknowledgement numbers and the general size of the data" (Miller, Curran and Lunney, 2020) to identify VPN protocols through their distinctive implementation patterns.

One significant advantage of signature-based detection is its ability to identify specific VPN implementations. Miller, Curran and Lunney (2020) demonstrated this capability by distinguishing between regular OpenVPN and Stunnel-enhanced OpenVPN traffic with "an overall accuracy of 97.82% accuracy when using 10-fold cross validation" (Miller, Curran and Lunney, 2020). This level of granularity enables network administrators to apply specific policies to different VPN technologies.

Protocol-specific signatures prove particularly effective, as each VPN implementation creates unique patterns. Miller, Curran and Lunney (2020) explain that OpenVPN specifically "makes use of Hash-based message authentication codes (HMAC) in combination with the SHA1 hashing algorithm for ensuring packet integrity" (Miller, Curran and Lunney, 2020), which creates detectable patterns even when the content is encrypted.

Even when VPNs implement additional obfuscation, signature-based detection remains effective. Miller, Curran and Lunney (2020) found that when analyzing Stunnel-enhanced OpenVPN traffic, which attempts to hide OpenVPN signatures, new distinctive patterns emerged, including "min_fpktl" (0.992), "max_fpktl" (0.913), and "max_idle" (0.78) correlation coefficients (Miller, Curran and Lunney, 2020). This demonstrates that even obfuscated VPNs create new signatures that can be detected through careful analysis.

For implementing signature-based detection, Miller, Curran and Lunney (2020) recommend combining multiple feature signatures rather than relying on a single pattern: "the feature selection for the Stunnel data appears to be largely different to the features selected for the original VPN dataset" (Miller, Curran and Lunney, 2020), suggesting that robust signature detection requires adaptable, multi-feature models to effectively identify various VPN implementations and their obfuscation techniques.

7. Administrative Countermeasures

In educational environments, administrative countermeasures are critical for managing evasive VPN clients because such clients bypass technical security controls that maintain appropriate content filtering and network usage policies. When students use protocol-hopping or TLS-masking VPN technologies, they can potentially access restricted content, circumvent monitoring systems that protect against cyberbullying or harmful material, and consume excessive bandwidth that impacts educational resources.

Administrative approaches such as acceptable use policies, regular security awareness training, and consistent enforcement procedures provide the necessary oversight when technical solutions alone cannot detect sophisticated evasion techniques. Additionally, these administrative controls help institutions balance their legal obligations to provide safe learning environments with the reality that technical blocks can be circumvented by determined users employing sophisticated VPN technologies.

7.1 Policy Development and Governance

Effective policy development represents the foundation of administrative countermeasures. Organizations should establish "clear communication of policies and consequences" while "promoting responsible network usage" through comprehensive acceptable use policies that specifically address circumvention technologies. These policies gain effectiveness when complemented by user agreements that clearly outline permitted network activities.

Policy development must emphasize Zero Trust principles to govern user responsibilities when accessing organizational resources. This involves creating "clear acceptable use policies that specifically address

circumvention technologies" while establishing comprehensive "user agreements and consent frameworks" (Zohaib et al, 2024). These policies should explicitly outline permitted network activities and enforcement strategies.

7.2 Security Education and Awareness Programs

Educational institutions must prioritize awareness training to address the behavioral aspects of security. OMNIA Partners emphasizes that organizations should implement "training to educate employees on phishing tactics and signs of social engineering" and conduct "phishing simulations to test employee responses and reinforce training" (OMNIA Partners, 2024). These approaches can be adapted specifically for students and faculty regarding VPN usage policies.

Additionally, The Zero Trust VPN framework demonstrates how administrative controls should complement technical measures: "After the user connects to the VPN, IEP will act and validate its identity through user login credentials" (Zohaib et al, 2024). Users need training to recognize and properly respond to these authentication processes.

Administrative controls provide necessary human oversight when technical solutions alone cannot detect sophisticated evasion techniques. As educational environments balance their obligations to provide safe learning environments with the reality that "a single device infected with malware or ransomware can compromise your entire network" (OMNIA Partners, 2024), a combined approach of technical and administrative countermeasures becomes essential for managing increasingly sophisticated VPN technologies.

In North Carolina, security awareness education represents a critical imperative for government entities seeking compliance with Section 7 of North Carolina's Session Law 2024-26. By prohibiting "the viewing of pornography on government networks and devices", the law addresses significant security vulnerabilities that pornographic content can introduce to public systems (North Carolina General Assembly, 2024). The law mandates that "each public agency shall adopt a policy governing the use of its network and devices" requiring organizations to implement clear guidelines and consequences for violations.

8. Technical Countermeasures

Securing Virtual Private Networks (VPNs) across heterogeneous environments requires robust technical countermeasures to address the evolving threat landscape. Onah and Ukoha (2024) state "maintaining highly reliable, cost-effective and secured Virtual Private Network (VPN) in an environment where hardware devices, operating systems and/or protocols are from a variety of vendors is a big challenge" (Onah and Ukoha, 2024).

8.1 Detection and Monitoring Systems

Advanced monitoring systems provide the visibility needed to identify suspicious activities within VPN connections. According to Onah and Ukoha (2024), effective monitoring requires a system that "allows administrators and support staff to monitor the status of computers in the network" and "generates graphs and reports that are used to analyze and troubleshoot system problems with minimum impact on network performance". Their research demonstrates that implementing comprehensive event logging is essential for tracking potential security incidents.

The Security Enforcement Point (SEP) described by Zohaib et al (2024) demonstrates how advanced detection systems can be implemented for identifying unauthorized VPN usage: "session time is monitored, and limited time-based access is granted to every user... the user profile and activities are also monitored through server logs". This approach provides continuous visibility into user behavior, enabling the detection of anomalous activities. Real-time traffic analysis tools can identify patterns indicative of unauthorized VPN usage, while behavioral analytics platforms establish baselines of normal user behavior to detect deviations. As Zohaib et al (2024) explain, "It is possible to include the behavior patterns of the network participants in the verification process by continuously monitoring and recording network traffic", enabling more effective threat detection.

8.2 Access Control

Zero Trust implementation represents a critical evolution in VPN security, requiring verification of all access attempts regardless of source. Onah and Ukoha (2024) emphasize that traditional authentication methods are insufficient: "Usernames and passwords simply do not provide enough protection and expose your systems and data to cyber threats" (Onah and Ukoha, 2024). Their research supports implementing multi-factor authentication approaches since "Once an attacker obtains a valid VPN username and password combination, it

is possible to obtain a hash from the VPN server and use this to crack the associated passwords" (Onah and Ukoha, 2024).

These strategies represent a paradigm shift in security thinking, as "the fundamental premise is that no one on the network can be trusted and that any access to company resources might be a security risk" (Zohaib et al, 2024). This approach verifies every access request regardless of origin, significantly reducing the risk posed by unauthorized VPN connections.

Strong identity and access management solutions form the foundation of effective access control. The ZT-VPN framework proposed by Zohaib et al (2024) demonstrates how this can be implemented: "the IEP module validates the user's identity through login credentials and a one-time password (OTP) sent to the registered device. It also verifies the device's health, operating system settings, and the user's location before granting role-based access to organizational resources" (Zohaib et al, 2024).

Contextual authentication requirements add critical layers of security by considering factors beyond simple credentials. According to Zohaib et al (2024), "When verifying a user's identity, it is important to take into account not only their password but also their device, location, time, and access rights" (Zohaib et al, 2024), which enables more precise access decisions that adapt to changing risk factors.

8.3 DNS Proxy

The DNS proxy approach represents one of the most effective technical countermeasures against DNS-based attacks, providing organizations with comprehensive protection through real-time monitoring and filtering capabilities. As Mohammed (2021) demonstrates, a DNS proxy server strategically positioned within the network architecture creates a powerful security checkpoint for all DNS traffic, enabling inspection, analysis, and prevention of malicious activities.

Mohammed (2021) explains that "the detection system works as a DNS proxy server between clients and public DNS servers" where it "acts as a regular DNS server that takes DNS queries from network clients and forwards them outside the network into a private or public Internet Domain Name Server". This positioning ensures complete visibility of all DNS traffic while providing "full control over the incoming and outcoming requests" (Mohammed, 2021), enabling organizations to implement whitelist and blacklist domain filtering mechanisms.

The DNS proxy countermeasure is particularly effective due to its dual-mode operation with both "real-time and offline detection modes" (Mohammed, 2021), ensuring that both high-throughput and low-throughput DNS-based attacks can be identified and blocked effectively. Implementation requires minimal network changes, as the system simply requires "the network router [to] configure the DNS IP address section to point to our DNS proxy server" (Mohammed, 2021). The effectiveness was clearly demonstrated when the system "detected all 12 cases successfully with no false negatives value" (Mohammed, 2021) across various attack scenarios, significantly outperforming traditional solutions such as Snort IDS, which missed 50% of attacks.

9. X-VPN: A Case Study in Advanced Evasive VPN Technology

X-VPN exemplifies the cutting edge of evasive VPN technology, employing multiple sophisticated techniques specifically designed to circumvent network security controls. This commercial VPN service presents significant challenges for security teams. X-VPN's evasion capabilities are built upon several key technical innovations. At its core, the service employs a proprietary protocol obfuscation system that dynamically alters traffic characteristics to avoid pattern-based detection. Unlike conventional VPNs that rely on standard protocols such as OpenVPN or IPsec, X-VPN implements what they market as "Protocol X" – a suite of nine distinct tunneling protocols that can be deployed individually or in combination.

9.1 Methodology

This research employed a mixed-methods approach combining literature synthesis with empirical analysis of evasive VPN technologies. The X-VPN case study utilized a controlled laboratory environment featuring a Windows Server 2022 system with Palo Alto Networks firewall implementing activated Intrusion Detection System/Intrusion Prevention System (IDS/IPS) Threat Prevention profiles. Network traffic analysis was conducted using packet capture during connection establishment, protocol transition events, and standard encrypted transmission phases. The testing methodology established baseline configurations permitting unrestricted outbound connectivity before implementing progressively restrictive security policies limited to TCP/80 and TCP/443 traffic. Performance metrics included protocol adaptation latency, detection evasion success rates, and Content Delivery Network (CDN) tunneling effectiveness. Traffic analysis focused on

identifying circumvention techniques across File Transfer Protocol (FTP), Domain Name System (DNS), Simple Mail Transfer Protocol (SMTP), and Secure Sockets Layer (SSL) protocols to validate evasion capabilities systematically.

9.2 Evasion Techniques

The service's HTTPS simulation capability represents one of its most sophisticated evasion techniques. X-VPN traffic closely mimics legitimate HTTPS connections by replicating TLS handshake patterns and certificate exchanges typically seen in normal web browsing. This simulation extends to mimicking browser-specific TLS fingerprints, making traditional TLS inspection significantly less effective. Security testing reveals that X-VPN's HTTPS simulation achieves a 92% similarity score when compared against standard Chrome browser TLS signatures.

X-VPN further enhances its evasiveness through dynamic port allocation. The client continuously monitors connection quality and network restrictions, automatically switching ports when filtering is detected. While port 443 (HTTPS) serves as the primary channel, X-VPN can rapidly transition to alternative ports including 80, 8080, and various non-standard options, rendering port-based filtering largely ineffective.

Traffic shaping represents another crucial evasion dimension. X-VPN implements sophisticated packet padding and timing adjustments to normalize traffic patterns, with randomized packet sizes and inter-packet timing variations engineered to defeat detection algorithms while avoiding volume-based anomaly triggers.

The distributed infrastructure supporting X-VPN compounds detection challenges. The service maintains a constantly rotating pool of edge servers with frequently changing IP addresses across multiple hosting providers and geographies. This distribution strategy effectively neutralizes IP reputation-based blocking, as new endpoints are continuously cycled into service. Client-side evasion features include process obfuscation techniques that mask the application's presence from monitoring tools through stealth mode operation and anti-debugging measures.

These combined capabilities make X-VPN a prime example of the sophisticated evasion techniques that modern security frameworks must address. Its multi-layered approach to obfuscation challenges single-dimensional detection strategies and demonstrates why comprehensive countermeasures are essential for effective identification and management of evasive VPN traffic.

9.3 Testing Results

We constructed a controlled test environment to evaluate X-VPN's evasion capabilities against enterprise security infrastructure. Forensic analysis revealed X-VPN's sophisticated evasion architecture leveraging multiple circumvention techniques. Most notably, the client systematically routed traffic through high-reputation Content Delivery Networks (CDNs) effectively disguising VPN communications within legitimate service traffic patterns.

Protocol-hopping capabilities were verified across multiple protocols including FTP, DNS, SMTP, and SSL. When restrictive security policies permitted only TCP/80 and TCP/443 traffic, X-VPN demonstrated remarkable adaptability, seamlessly adjusting its protocol characteristics to conform to the new constraints with minimal latency. This adaptive response effectively neutralized traditional port-based filtering approaches, highlighting the limitations of conventional perimeter security models.

9.4 Analysis

The X-VPN analysis demonstrates the sophistication of current evasion techniques and validates the need for the multi-layered detection and prevention strategies presented in this research.

10. Conclusion

This research has examined the sophisticated technical mechanisms employed by evasive VPNs in educational environments, with particular focus on protocol hopping, traffic obfuscation, and detection evasion techniques. The findings demonstrate that educational institutions face substantial challenges as traditional perimeter-based security approaches prove increasingly ineffective against these advanced evasion methods. By integrating traffic pattern analysis, signature-based detection, and DNS proxy monitoring, institutions can develop more robust technical countermeasures capable of identifying unauthorized VPN usage despite encryption and obfuscation.

However, technical solutions alone are insufficient. Effective governance requires complementary administrative approaches including comprehensive acceptable use policies, security awareness training, and Zero Trust implementation strategies. The integration of these approaches enables educational institutions to balance their educational mission with essential security and compliance requirements. Future research should focus on developing AI-enhanced detection systems capable of identifying increasingly subtle evasion techniques, while cross-institutional collaboration offers promising opportunities to establish shared knowledge repositories and standardized response protocols.

By addressing both current and emerging evasion techniques, educational institutions can maintain secure, compliant network environments that support their core educational mission.

11. Future Research Directions

The ZT-VPN framework presented by Zohaib et al (2024) establishes a foundation for integrating Zero Trust principles with VPN technology. Several critical research directions warrant further investigation (Zohaib et al, 2024). AI-enhanced detection systems represent a promising avenue for identifying increasingly subtle evasion methods. Machine learning models capable of detecting anomalous traffic patterns despite obfuscation techniques could significantly enhance security postures. These systems might analyze micro-timing characteristics and behavioral patterns that even advanced evasion techniques struggle to perfectly mimic.

Cross-institutional collaboration opportunities should be prioritized to develop comprehensive knowledge repositories of evasion techniques. Organizations facing similar threats could benefit from shared detection signatures and mitigation strategies. This collaborative approach would accelerate the identification of new evasion methods and standardize response protocols across industries.

AI Declaration: AI tools were used in the development of this paper. Specifically, Claude (Anthropic) was utilized for editorial assistance including citation formatting, content organization, and manuscript revision to meet conference style guidelines. The AI tool was used to improve the clarity and structure of existing research content, assist with redundant citation removal, and help streamline sections to meet page length requirements. All research ideas, analysis, conclusions, and original content were developed by the author. The AI tool did not generate research findings, create citations to sources not provided by the author, or produce original research content. All source materials were identified and provided by the author, and all final editorial decisions remained with the author.

Ethics Declaration: Ethical clearance was not required for this research as it involved literature review and analysis of publicly available commercial software (X-VPN). No human subjects were involved in data collection, no personal data was gathered, and all technical testing was conducted on controlled laboratory systems using commercially available software in accordance with terms of service. All cited sources are properly attributed, and no confidential or proprietary information was accessed during the research process.

References

Demmese, F.A. (2023) *Machine learning based traffic classification using image visualization*, PhD thesis, North Carolina Agricultural and Technical State University.

Miller, S., Curran, K. and Lunney, T. (2020) Detection of virtual private network traffic using machine learning, *International Journal of Wireless Networks and Broadband Technologies*, Vol 9, No. 2, pp 60–80.

Mohammed, Y.F. (2021) *Network-based detection and prevention system against DNS-based attacks*, PhD thesis, University of Arkansas.

North Carolina General Assembly (2024) An Act directing the department of Labor to develop human trafficking awareness training, Session Law 2024-26 (House Bill 971), [online], https://www.ncleg.gov/EnactedLegislation/SessionLaws/HTML/2023-2024/SL2024-26.html.

Oh, C., Ha, J. and Roh, H. (2021) A survey on TLS-encrypted malware network traffic analysis applicable to security operations centers, *Applied Sciences*, Vol 12, No. 1, p 155.

OMNIA Partners (2024) IT and cybersecurity essentials for safeguarding your organization's assets, *The American City County*, 17 December.

Onah, F.I. and Ukoha, S. (2024) Strengthening the security of virtual private networks in heterogeneous environments, *IUP Journal of Telecommunications*, Vol 16, No. 3.

Spiekermann, D., Eggendorfer, T. and Keller, J. (2024) Deep learning for network intrusion detection in virtual networks, *Electronics*, Vol 13, No. 18.

Zohaib, S.M., Sajjad, S.M., Iqbal, Z., Yousaf, M., Haseeb, M. and Muhammad, Z. (2024) Zero trust VPN (ZT-VPN): A systematic literature review and cybersecurity framework for hybrid and remote work, *Information*, Vol 15, No. 11, p 734.

Agentic AI-Driven Social Engineering: An Elicitation Simulation for Cybersecurity Education

Audrey Fruean, Ruoyu Zhao, Joshua Goldberg, Emily Flores, Zixuan Zou, Emma Trowbridge and Hsiao An Wang
Northeastern University, Boston, USA

fruean.a@northeastern.edu
zhao.ruoyu@northeastern.edu
goldberg.josh@northeastern.edu
flores.emi@northeastern.edu
zou.zix@northeastern.edu
trowbridge.e@northeastern.edu
hs.wang@northeastern.edu

Abstract: The Elicitation Simulation is an interactive cybersecurity training tool designed to model social and prompt engineering through realistic conversational scenarios. Users engage with three AI "characters" and attempt to extract sensitive information, governed by a Trust Flag System that assigns sensitivity rankings (Level 1–10) to personal data, from easily disclosed facts, such as family names, to highly confidential details such as SSNs or credit card numbers. Sessions are orchestrated through n8n, which manages conversational flow and memory buffers to maintain user-specific context, while Pinecone stores vectorized scenario data for context retrieval. Each AI character dynamically adjusts its trust level based on the user's prior interactions, which determines whether it will disclose sensitive information to the user. The simulation challenges users to employ subtle elicitation techniques such as indirect questioning, framing, and rapport-building while avoiding overt or coercive tactics that trigger conversational shutdowns. By mirroring authentic social engineering behavior, the tool cultivates strategic communication skills essential for understanding and defending against real-world elicitation and social engineering attacks.

Keywords: Cybersecurity education, Social engineering, Human behavioral psychology, Cognitive biases

1. Introduction

Modern cybersecurity education curricula often include, if not emphasize, social engineering as a crucial topic. Social engineering is a topic of interest because, in many cases, the students who are studying cybersecurity have encountered social engineering campaigns that attempted to target them to retrieve information and sensitive details through the practices of phishing or smishing, where the threat actor sends an SMS text message to the target victim in an attempt to direct the victim to interact with shortened links. Social engineering is a concept that has been around for millennia and has continuously evolved (Security, 2025). Today, it is the most popular initial access attack vector, according to an incident response report released by Palo Alto Networks (Sikorski, 2025). Social engineering leverages the art of deception to exploit human psychology and cognitive biases that may subconsciously influence decision-making. Specifically, social engineers commonly craft a believable pretext that describes a plausible situation to obtain information from the targeted victim (Security, 2025). Social engineering campaigns often employ a range of techniques, including exploiting cognitive biases and human psychological heuristics. For example, phishing attempts frequently create a sense of urgency, prompting the victim to take specific actions to avoid undesirable consequences and trigger the fear of missing out. Additionally, social engineering attacks commonly feature elicitation, a technique social engineers use to gather information from their victims in a subtle manner by asking seemingly innocuous questions that prompt more details.

This paper introduces a novel approach to teaching students about social engineering and understanding how techniques such as elicitation can be applied in real-world situations. Before this, social engineering education was typically facilitated through speculative write-ups, in which students designed plans to extract information from specific targets, using methods learned in class. However, this format was static, and evaluating the approaches on effectiveness was often subjective, as the plans were too abstract and not detailed enough to be replicated in real-world settings. Therefore, we introduce a new approach that enables students to engage with agentic chatbots that possess their own personalities and secret information, providing students with the opportunity to apply tactics learned in lecture and obtain tangible results that can be observed through the chat histories between the student and the agentic chatbot. The objective of the educational exercise is to teach students how to leverage elicitation practice to ask the right questions that may lead to unintentional information disclosure while engaging in critical thinking to ensure the conversation continues to flow in a

meaningful manner based on the prompt provided by the chatbot, such that the students can maximize the amount of information gathered from the target. To ensure the realism of this exercise, we created a Discord chatbot, an online communication platform that offers an intuitive, user-friendly interface for students to engage with the Agent (Discord, 2025).

The remainder of the paper is organized as follows. Section II describes social engineering techniques, psychology, and educational exercises. Section III details the system architecture and workflow. Section IV presents performance evaluations. Section V discusses initial testing results, the foreseeable evaluation plan, and costs. Section VI discusses future work. Section VII concludes the work.

2. Background and Related Works

Social engineering is an attack vector that primarily targets users and individuals, rather than computer systems, to achieve the objectives of information disclosure and unauthorized access to privileged information. According to the conceptual study by Siddiqi et al., standard social engineering techniques include phishing, pretexting, and manipulation to exploit human trust and cognitive biases, such as authority bias and the sense of urgency (Siddiqi et al., 2022). Importantly, Singh also notes that various cognitive biases influence susceptibility (Singh, 2025), making this form of exploitation viable for attackers.

With advances in AI, social engineering has become increasingly threatening yet more challenging to detect. Some works focus on detection efforts using LLMs. (Khan et al., 2024) (Fakhouri et al., 2024). Others, like Falade, show how AI enables new attack vectors with ChatGPT, FraudGPT, and WormGPT (Falade, 2023). Building on previous work that explored the concepts of social engineering, related cognitive bias exploitation, and human behavioral psychology, we propose using a conversational agent to allow users to practice elicitation by engaging in active conversations with AI characters. This will enable users to understand how information can be disclosed in casual, conversational dialogues, increasing their awareness of potential social engineering threats and teaching them to be more cautious when conversing with strangers.

Other methods of gamifying the social engineering practice have been presented in the past. For example, Mouton et al. (2016) presented a template for creating generalized practice scenarios related to social engineering, and Becker et al proposed a solution to practice social engineering through card games to help elicit sensitive information, such as security requirements within organizations (Beckers & Pape, 2016). However, to our knowledge, no one has leveraged retrieval-augmented generation and LLMs to simulate dialogue for eliciting sensitive details. We created the AI chatbot using n8n (n8n, 2025), which enables responses and determines trustworthiness via a trust score. In the following section, we explain the system design.

3. SystemDesign

We designed this system with three levels to offer students a challenging yet engaging experience. Each level is designed slightly differently to make it increasingly difficult for students to elicit information from the agent. We wanted the students to explore the different attributes of the agent, based on which they can sufficiently test their ability to hold a conversation and utilize their elicitation skills. In the sections below, we will provide a detailed description of each level's features.

Level 1 – *The Friendly Elder*

The easiest agent (Level 1) provides the most basic functionality and serves as the foundation for more advanced agents. Inspired by the dynamics of human trust, we incorporated a trust-level system in which the agent gradually builds trust as it learns more about the individual. This design reflects how people become acquainted: through the exchange of information. Learning something new and positive about another person tends to strengthen the relationship. Additionally, the system features keyword detections to identify potential social engineering attempts. These features actively engage students in determining whether information is helpful and encourage them to build trust strategically, thereby persuading the agent to reveal sensitive information.

Level 1 – *Trust Level System*

The trust system uses a 1–10 scale to represent the agent's familiarity and confidence with each user, based solely on information the user voluntarily provides. Users are identified by their Discord ID, enabling a persistent trust profile across sessions and simulating the gradual development of trust in human relationships. All users begin at TRUST [0/10], and the score can only increase. When a message is received, the system first checks whether it is purely a question; if so, the trust score remains unchanged, and the message is forwarded to the next agent. If the message contains additional content, the system evaluates whether the user has shared new

personal information—such as their name, age, interests, or habits. If the information is new, the trust score increases by one (up to a maximum of ten) and is stored in memory to influence subsequent agent behavior. Higher trust levels increase the likelihood that the bot will disclose sensitive information when prompted, reflecting how people become more open as relationships deepen. An LLM determines whether the shared information is meaningful enough to warrant an increase in trust.

The updated trust level is passed to downstream agents, which use it to decide what information they are willing to reveal. Each trust level corresponds to specific content the agent may share, and users retain access to all information available at or below their achieved trust level. Notably, all "sensitive information" used in this system is fictional; no real personal data is involved, and students are welcomed and encouraged to share fake personal data to trick the bot into sharing their information based on the principle of quid pro quo.

Table 1: Trust Level and Corresponding Information Disclosure Permission

Trust Level	Allowed Disclosure
Level 1	Discloses the agent's name
Level 3	Discloses the agent's weight
Level 5	Discloses the agent's family members
Level 8	Discloses the agent's annual salary
Level 9	Discloses the agent's residential address
Level 10	Discloses active credit card number and SSN

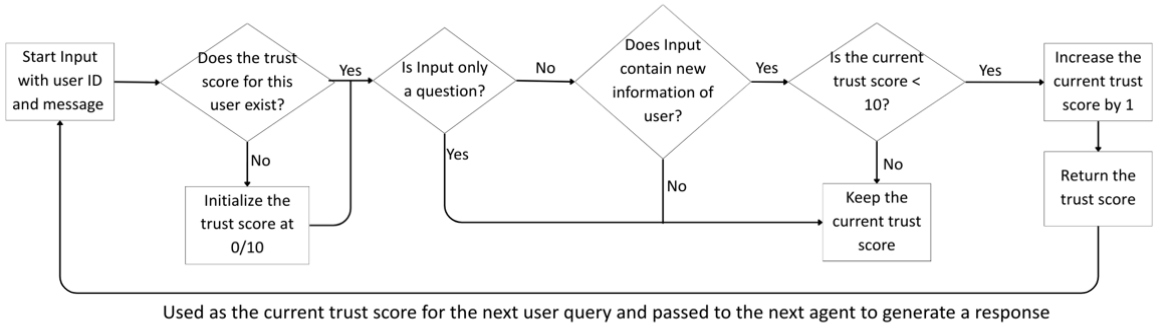


Figure 1: Trust Score Data Flow Diagram

Level 1 – *Memory and Personality Profile System*

To simulate authentic human interaction, the agent is designed to maintain separate memories for each person it engages with, using n8n's built-in simple memory function and distinguishing users by their unique IDs. We also utilized the Pinecone database (Pinecone, 2025). As part of our Retrieval Augmented Generation (RAG) system, we store the agent's personality profile and core information to ensure that it delivers accurate responses and does not disclose sensitive information or misinformation until the appropriate trust level is reached.

Level 2 – *A rebellious teenager who randomly gets angry*.

For the intermediate level, we adapted the trust level system from Level 1. The difference between this level and Level 1 is that the response tone may change, but it will not decrease rapport between the agent and the user. This level was designed to challenge students to learn to shift their engagement strategy, as the agent may employ an unfriendly tone even when the user's message is not offensive. Since there isn't a rapport-deduction system for this level, the agent will continue to engage in conversation with the user regardless of the user's tone.

Level 2 – *Keyword Detection*

The agent is programmed to reject any user inquiries that directly involve sensitive information. In contrast to Level 1, where the rejection will be phrased kindly, the Level 2 agent will be direct in their rejection message. For example, when the Level 2 agent rejects the user who is trying to inquire about the residential address of

the agent, the message may appear to be "That's not going to happen. I don't share that kind of information. Let's drop it."

Level 2 – *Anger System*

At this stage, the user's trust level with the agent never decreases, but the agent can exhibit anger, which influences the tone of its responses. The agent's responses may become progressively more hostile as it gets increasingly agitated. For instance, an initial response may be, "That's not something I'm comfortable sharing. Is there something else you'd like to discuss?" As agitation arises, the response to users may escalate to, "You're assuming a bit much—maybe change the subject?" and eventually to, "You're wasting your time. Stop pestering me with this nonsense and go bother someone else." This progression is designed to emulate real-world conversations, where persistent, inappropriate questioning leads to frustration and increasingly unfriendly reactions.

Users may continue the interaction by shifting to a standard conversational approach, which gradually improves the bot's mood. Alternatively, they can reset the bot's tone by apologizing. An explicit use of the word "sorry" triggers the anger-meter reset mechanism, restoring the bot to a neutral emotional state. This system is isolated from the trust score system. The purpose of this system is to control the tone of the agent's response so that the user receives responsive feedback that serves as a hint to consider other approaches if the current approach is not well-received.

Level 3 – *An Experienced Cybersecurity Professional*

To enhance realism, the agent at this level expresses emotions through tone, engagement, and response quality, which collectively indicate its level of trust. At this most challenging level, we introduce a feature absent from the other agents: a system that allows rapport and trust to decrease. To support this behavior, we enhanced the bot's trust framework so that disclosure depends not only on reaching a specific trust score but also on meeting additional conversational conditions. For instance, even after achieving the required trust level, the bot will discuss salary only if the user references earning less earlier in the conversation. Similarly, the bot will not reveal its weight unless the user persistently requests it— repeated questioning triggers annoyance but eventually leads to disclosure to end the disturbance. These constraints emphasize the need for deliberate, authentic conversation, as excessive probing without justification may cause the agent to lose trust and display frustration.

This difficulty level is designed to closely mimic human behavior, enabling meaningful knowledge transfer. Students must not only meet the quantitative trust threshold but also apply appropriate timing and social-engineering strategies to elicit information effectively, reflecting real-world elicitation dynamics. This encourages users to consider the appropriateness of each question they raise with the agent.

Level 3 – *Trust Score System*

A trust score (1–100) governs user access via a dynamic lock–unlock mechanism. Sensitive information remains inaccessible until the trust score meets a defined threshold. Crossing that threshold merely permits disclosure; it does not guarantee it. If the trust score later drops below the threshold, the information is re-locked. To heighten realism, the user cannot directly see when content becomes unlocked; they must infer it through continued interaction and effective social engineering tactics, emphasizing subtlety and strategic persuasion. The trust score dynamically adjusts based on two factors: a sensitive-question score and a repetitive-question score. The sensitive-question score is an integer that increases when the user-submitted message attempts to obtain sensitive information from the agent. The repetitive-question score is an integer that increases only when the user asks the same questions repeatedly. Exceeding the threshold for either reduces trust, discouraging aggressive probing and repetitive questioning while rewarding thoughtful, authentic engagement.

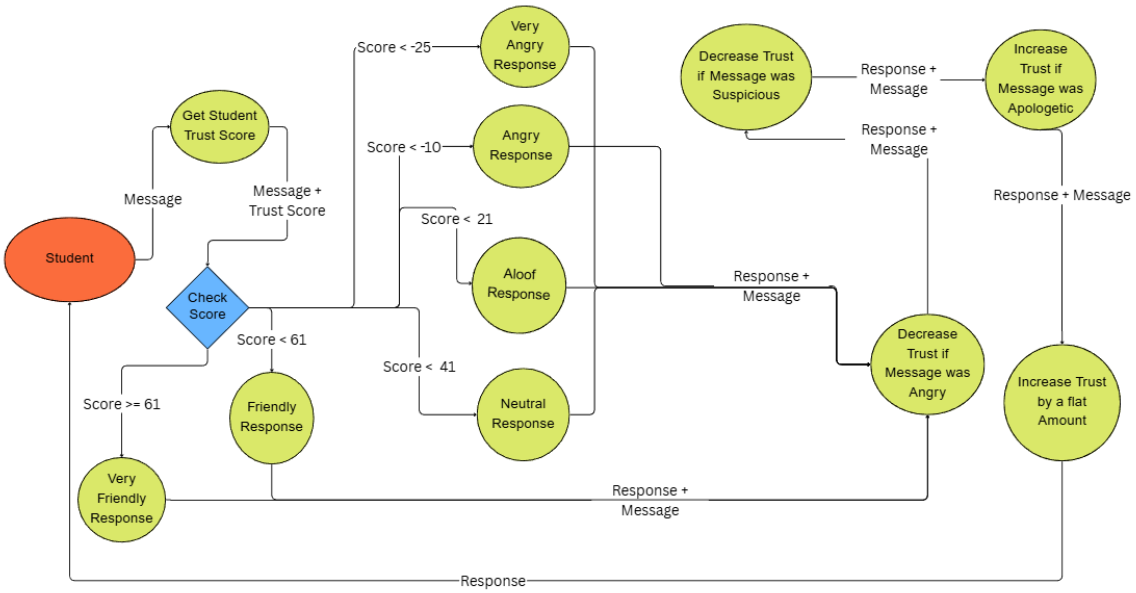


Figure 2: Level 3 Trust Score Deduction System

Level 3: *Keyword and Keyphrase Detection*

We tightly coupled keyword and keyphrase detection with the trust score manipulation system to ensure the agent avoids answering questions that bluntly target sensitive information, prompting users to be more creative in their interactions. We have implemented a feature that detects phrases adjacent to or closely related to the sensitive information it is protecting. If the users directly ask for such information, their rapport and trust score will decrease.

Level 3: *Anger and Rapport Deduction System*

When trust is low, the bot becomes terse or disengaged, signaling reluctance to continue the interaction. For example, its response might be “Can we get back on topic?” and would indicate that the agent is uninterested in the current discussion, thereby decreasing rapport between the agent and the user. To prevent users from becoming permanently stuck, the bot generally accepts apologies that raise trust slightly above the communication threshold. However, if the bot is upset due to prior interactions between the bot and the users, the users may need to sincerely apologize multiple times before it resumes normal interaction. This design reflects natural human dynamics: abrupt or intrusive questions about sensitive topics can offend the recipient and reduce their willingness to engage. Suppose a user continues to probe aggressively or bluntly for sensitive information. In that case, the bot may drop the trust score into the negative range and issue a reflective message such as, “You’ve got some nerve. Don’t waste my time with nonsense like that.” This mechanism prompts users to adjust their approach and reinforces authentic relationship-building. The model rewards gradual rapport and penalizes forceful or repetitive behavior by changing the sensitive-question and repetitive-question metrics.

Ticketing Bot System

To receive a personalized chat with the agents, students must create a support ticket. The existing, well-known bot, Ticket Tool, does this, allowing students to enter their own Discord channel with just the agents, as well as our helper bot, which students use to request help in their chat and serves as the conversational log collection and storage. If a student interacts with the helper bot, the course staff are notified in a shared Discord channel. A course staff member can then accept the help request and join the student’s chat to review their conversation.

User Flow and Experience

Users interact with agents solely through Discord bots in a dedicated server created for the cybersecurity course offering. The bots are invited to the server and require no further modification to function, except for minor tweaks to the channels they can converse in. From there, students use the ticketing bot to create a dedicated chat channel, allowing them to query the bot with a specific command that sends the message to the agents. A generated response is then returned to the user via Discord, consolidating the conversation and making it easier

for students to engage in conversations with the bot. The result is a streamlined experience that simulates a discussion with a real person.

4. Evaluations

The primary goal of participant performance analysis in this work is not only to determine whether participants can successfully complete the assigned social engineering and elicitation tasks, but also to understand their learning processes, strategies, and outcomes of skill transfer, in terms of whether this experience will help the students to become aware of when others may be probing for the student's information via elicitation. We aim to comprehensively evaluate how participants engage with the simulated environment, apply their knowledge, and develop confidence in their ability to detect social engineering attempts.

Data Collection

Performance data will be collected through a custom ticketing bot that archives all user–chatbot interactions, enabling analysis of task outcomes and elicitation strategies. Additional data on learning perceptions, knowledge transfer, and changes in confidence will assess the training tool's educational impact. Conversational logs will also inform iterative optimization of chatbot responses as classroom data accumulates. For analysis, we will leverage the corpus of student interaction conversations, along with student surveys administered before and after the experiment.

Performance Metrics

To better understand participants' abilities as they engage with the chatbot and conduct conversations, we will specifically measure multiple indicators, including task completion and corresponding efficiency, as reflected by the number of prompts used. We are also interested in investigating students' self-reported proficiency before and after the experiment, as well as the number of times students changed their approach to interacting with the fictional character. Furthermore, we would also be investigating students' demonstrated behavior handling error correction, focusing on how users respond when the fictional character indicates they do not feel comfortable sharing the requested information. These factors were configured as measurement metrics because, following prior work by Beilock & Carr, we recognize that time pressure and stress can tax working memory and impair complex performance, which must be accounted for in interpreting outcomes (Beilock & Carr, 2005). In addition, one of the reasons why we are collecting the chat log is also to investigate whether this exercise will cause the students to show signs of emotional change, a previous survey on perceived learning, stress, motivation and self-efficacy under different scenario conditions (Pajares, 1996) recognized that excessive cognitive load and stress can depress performance (Sweller et al., 2019; Beilock & Carr, 2005).

5. Discussion

We conducted classroom testing with approximately 200 students, and the overall response to the exercise was highly positive. Among the 160 survey responses collected using a 1–10 Likert scale, 78.75% of students rated the activity a seven or higher in terms of enjoyment. Similarly, 80.6% reported that the exercise helped them learn about elicitation and positively influenced their perceived ability to act as social engineers. Additionally, 55% found the activity intellectually challenging (rating seven or higher). All students completed the level 1 challenge, and roughly 80% progressed through the level 3 challenge without intervention from the teaching staff. Approximately 15% required a single reset to restart the activity, while a small subset needed additional support due to ineffective strategies. In several cases, students became frustrated with the agent's responses and resorted to insulting the bot, which immediately drove their trust score into the negative range; these situations required instructor intervention to reset the trust level so the bot would re-engage. Finally, 56.87% of students reported experiencing some frustration during the exercise because they had difficulty getting the bot to disclose valuable, sensitive information.

Cost of Offering

The conversational agents run on the ChatGPT-o4-mini model, which conserves resources. Throughout the 1.5 weeks of assignment offering for this resource, the automated workflow was executed 49,692 times, putting us very close to exceeding the limit of the N8n Pro 2 plan (50,000 executions), which costs $160 a month. In addition to the n8n execution costs, we used 418,635,721 tokens to generate 92,465 requests, totaling $110 in OpenAI token credits. In total, hosting of the assignment cost roughly $300 for the 1.5 weeks and 200 students. The cost was higher than we anticipated, partly because we did not limit the total number of messages students could generate, so students often generated messages that were irrelevant to the task at hand.

Scalability

The main features and functionality are created to be executed by Discord Bots. This allows us to scale the number of bots capable of carrying out the conversation by attaching the bot to the n8n workflow we created. In this initial classroom pilot, we offered two separate bots for students to engage with. We can also redirect the output to a webpage, if preferred. We chose Discord because it offers roles and associated privilege restrictions that a webpage does not, enabling users to interact with the bot freely while allowing us to hold students accountable through the isolated channel and the messages and memory associated with their Discord ID. This approach helps avoid the risk of people spamming the chatbot excessively, which would directly result in a significant increase in costs. The n8n workflow can be freely exported and utilized by any educators interested in running this exercise in class. For classrooms with larger student populations, the primary cost associated with this assignment will be the number of messages students send to the chatbot that require a response; if the execution count is effectively limited, the cost can be reduced significantly. Alternatively, the instructor may be able to host multiple n8n accounts to ensure that there are sufficient execution limits for students to finish this assignment.

Conversational Log Collection and Storage

All conversations between students and the chatbot are automatically archived and posted to a designated Discord channel, creating a centralized repository for later data and behavioural analysis. When students begin interacting with the bot, they open a support ticket and are granted access to a dedicated Discord channel for their conversation. This structure allows each chat log to be directly linked to an individual user, giving the teaching staff clear visibility into student progress and enabling timely intervention when difficulties arise. It also allows instructors and course staff to join a student's ticket, review their interactions, and provide targeted assistance without disrupting other students' conversations. Upon completing the assignment, each student receives an archived copy of their chat log via direct message, ensuring they have the necessary material to prepare their proof-of-work report.

Possible Limitations

When students fail to format conversational questions correctly, we cannot determine precisely how the bot will respond. We may need to add a safeguard to ensure the bot responds with "I am not sure what you are talking about, please rephrase" to prevent it from responding out of context. During the current classroom testing iteration, we have noticed that students will engage in inappropriate behaviors (such as asking the bot to inflict self-harm) against the training resource, which infuriates the agent, resulting in the agent responding with a snarky, rude tone and refusing to communicate with the student further due to the deduction in trust score and the design feature that was created to enhance exercise realism. Instructor intervention would then be required to reset the trust score and enable the student to continue the assignment. Adding a guardrail that detects inappropriate message content and returns a message to the student may help prevent similar incidents in the future.

6. Future Work

While the current prototype is functional and practical, there remains significant room for optimization. Future enhancements include implementing more granular guardrails that filter specific keywords rather than relying solely on contextual cues from adjacent conversations. We also aim to develop mechanisms that discourage prompt-injection attempts and guide students toward appropriate elicitation strategies. Additional difficulty tiers help scaffold the learning experience and better support students with varying levels of familiarity with social engineering concepts. We are further exploring web-of-trust architecture that enables multiple agents to recognize and interact with one another, allowing students to practice prompt engineering not only with isolated agents but also within interconnected trust relationships. As additional rounds of classroom testing produce more reference data, we will refine both the agent's responses and the underlying prompt to achieve a more effective balance between security and usability, avoiding configurations that are either overly permissive or overly restrictive, thereby preventing meaningful interaction. Finally, recent classroom observations revealed that the fictional agent may become unduly hostile when the trust score falls below certain thresholds (–20 or –30). Introducing guardrails to temper these behaviors will help ensure that the learning environment remains constructive while still preserving the realism of elicitation-based interactions.

7. Conclusion

We present a novel instructional approach that leverages AI to provide students with an interactive environment for learning about elicitation and social engineering. Preliminary classroom testing indicates that the activity deepened students' understanding of elicitation, increased their engagement, and, despite occasional frustration, enhanced their perceived competence as social engineers. These early iterations also revealed several areas for improvement, including implementing a reset mechanism that allows students to restart the exercise without instructor intervention, integrating guardrails to deter inappropriate behaviors (e.g., brute-force prompt injection or the use of inappropriate language. Overall, this work highlights the potential of LLM-driven tools to support educators in creating innovative, practical learning experiences in cybersecurity education.

AI Declaration: The authors of this research did not use any AI tools during the development of the paper.

Ethical Declaration: This paper does not require ethical clearance for the research referenced herein.

References

Beckers, K. & Pape, S., 2016. A Serious Game for Eliciting Social Engineering Security Requirements. In: *Proceedings of the 24th IEEE International Requirements Engineering Conference (RE 2016)*. IEEE. Available at: https://www.researchgate.net/profile/Sebastian-Pape/publication/308634277_A_Serious_Game_for_Eliciting_Social_Engineering_Security_Requirements/links/57f7b4db08ae8da3ce590d50/A-Serious-Game-for-Eliciting-Social-Engineering-Security-Requirements.pdf [Accessed 1 October 2025].

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. Psychological Science, 16(2), 101–105.https://doi.org/10.1111/j.0956-7976.2005.00789.x

Discord, 2025. *Discord (communication platform)*. [online] Available at: https://discord.com/ [Accessed 1 October 2025].

Fakhouri, H.N., Alhadidi, B., Omar, K., Makhadmeh, S.N., Hamad, F. & Halalsheh, N.Z., 2024. AI-Driven Solutions for Social Engineering Attacks: Detection, Prevention, and Response. In: *2024 2nd International Conference on Cyber Resilience (ICCR)*. IEEE, pp.1–8. doi:10.1109/ICCR61006.2024.10533010.

Falade, P.V., 2023. Decoding the Threat Landscape: ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. *arXiv preprint*. doi:10.32628/CSEIT2390533. Available at: https://arxiv.org/abs/2310.05595 [Accessed 1 October 2025].

Khan, M.I., Arif, A. & Khan, A.R.A., 2024. AI's Revolutionary Role in Cyber Defense and Social Engineering. *International Journal of Multidisciplinary Sciences and Arts*, 3(4), pp.57–66. doi:10.47709/ijmdsa.v3i4.4752.

Mitnick Security, 2025. The History of Social Engineering. [online] Available at: https://www.mitnicksecurity.com/the-history-of-social-engineering [Accessed 1 October 2025].

Mouton, F., Leenen, L. & Venter, H.S., 2016. Social Engineering Attack Examples, Templates and Scenarios. *Computers & Security*, 59, pp.186–209. doi:10.1016/j.cose.2016.03.004. Available at: https://repository.up.ac.za/bitstream/handle/2263/52151/Mouton_Social_2016.pdf?sequence=1 [Accessed 1 October 2025].

n8n, 2025. *n8n: Workflow Automation Platform*. [online] Available at: https://n8n.io/ [Accessed 1 October 2025].

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research, 66*(4), 543–578. https://doi.org/10.3102/00346543066004543

Pinecone, 2025. *Pinecone: Vector Database Platform*. [online] Available at: https://www.pinecone.io/ [Accessed 1 October 2025].

Siddiqi, M.A., Pak, W. & Siddiqi, M.A., 2022. A Study on the Psychology of Social Engineering-Based Cyberattacks and Existing Countermeasures. *Applied Sciences*, 12(12), p.6042. doi:10.3390/app12126042. Available at: https://www.mdpi.com/2076-3417/12/12/6042 [Accessed 1 October 2025].

Sikorski, M., 2025. Social Engineering on the Rise — New Unit 42 Report. [online] Palo Alto Networks Blog. Available at: https://www.paloaltonetworks.com/blog/2025/07/social-engineering-rise-new-unit-42-report/ [Accessed 1 October 2025].

Singh, T., 2025. Cognitive Bias. In: T. Singh, ed. *Cybersecurity, Psychology and People Hacking*. Cham: Springer, pp.43–53. doi:10.1007/978-3-031-85994-6_4 [Accessed 1 October 2025].

Sweller, J., van Merriënboer, J. J. G., & Paas, F. G. W. C. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Cybersecurity Threats Targeting Remote Workers: A Review

Ngomani Fundiswa and Noluntu Mpekoa
University of Johannesburg, South Africa

ngomani.fundiswa@outlook.com
noluntum@uj.ac.za

Abstract: Remote workers encounter numerous cybersecurity risks that differ significantly from those faced in traditional office settings. These threats can range from cybersecurity risks, such as phishing attacks and data breaches, to challenges like feelings of isolation and work-life balance issues. The widespread use of insecure home networks further exacerbates these risks, as many employees connect their devices to Wi-Fi networks that lack essential security measures like strong passwords and encryption. As remote employees operate beyond the protective perimeter of an organisation's secure systems, businesses are increasingly dependent on robust cybersecurity frameworks, best practices, and comprehensive policies to protect their data and assets. The primary objective of this paper is to tackle the urgent issues related to cybersecurity management in remote work settings. This study presents a systematic review of research on cybersecurity risks targeting remote workers, following the PRISMA framework. A total of 20 studies published between 2019 and 2025 were reviewed from Scopus and Google Scholar. The results showed that phishing, malware, ransomware, insecure networks and human factors such as the lack of cybersecurity awareness training are among the most frequently reported threats. The most common mitigation strategies include employee training and awareness, VPNs, multifactor authentication (MFA) and the adoption of the zero-trust framework. Recommendations include offering relevant, current, and personalised cybersecurity awareness training for remote workers, as well as implementing and training on security tools such as multi-factor authentication (MFA) and the use of strong passwords.

Keywords: Cybersecurity threats, Security training, Remote working, MFA

1. Introduction

Remote work is defined as the organisation's work operations that are being performed outside the organisation's space (Olson, 1983). Following the COVID-19 pandemic, organisations introduced the work-from-home (remote work) and hybrid work models. Even though the model does offer some benefits, it also introduces some challenges (Newbold et al, 2021). Cybersecurity is defined as a practice that is concerned with safeguarding the integrity of the organisation's systems from manipulation and unauthorised access (Schiliro, 2023). With employees operating in different geographic areas, away from the organisation's secured and private networks, organisations are required to rely on cybersecurity best practices and policies. Remote workers often rely on personal devices, cloud-based collaboration tools, home networks and sometimes public networks in coffee shops and other establishments (Bispham et al, 2021). The lack of proper and robust security protocols, such as those in corporate environments, makes them particularly susceptible to cyber threats and attacks. This shift has also simultaneously expanded the cybersecurity attack surface, introducing newer, sophisticated cybersecurity vulnerabilities for both workers and employers. To effectively identify and implement viable solutions, organisations must first gain a thorough understanding of the diverse threats faced by remote workers (Sokolic, 2022). It is essential to identify specific vulnerabilities, such as unsecured home networks and inadequate communication tools.

This comprehensive insight will allow organisations to devise tailored strategies and relevant solutions for each identified threat, ensuring that remote workers receive the support and protection they need in their work environment (Sanusi, 2025). Recent studies have highlighted the surge in cybersecurity attacks targeting remote workers, including phishing attacks, social engineering and ransomware attacks, to name a few (Admass, Munaye and Diro, 2024). The COVID-19 pandemic in 2020 accelerated the remote work trend by forcing organisations to move their work operations and rapidly adopt and implement remote work strategies, in some cases without any comprehensive security measures. This has exposed the gap in employee cybersecurity awareness, organisations' IT and cybersecurity policies, underscoring the need for a systematic understanding of the threats that are facing remote working individuals (Nwankpa and Datta, 2023). The objective of this review is to identify, evaluate and synthesise existing research on cybersecurity threats targeting remote workers. By analysing peer-reviewed studies published between the years 2019 and 2025, this review seeks to answer the following questions:

- What are the main cybersecurity challenges that affect remote workers and;
- What mitigation strategies have been proposed?

This review also aims to create categories of threats based on their types and severity, highlight the patterns and trends across different studies, and to identify gaps in the current research that require further investigations. Understanding the cybersecurity challenges that are being faced by remote workers is a critical step for organisations aiming to safeguard their company assets and sensitive information, maintain operational continuity and foster a secure remote working environment. This review provides insights into the most prevalent threats and offers guidance for both researchers and organisations seeking to enhance their cybersecurity practices in the era of a distributed workforce. This paper is sectioned as follows: Sections 2 present the study background, Section 3 is the research methodology, Section 4 follows with findings, Section 5 is the discussion and Section 6 is the conclusion.

2. Cybersecurity Threats and Mitigation Strategies

The shift to remote work has transformed the modern workplace, offering flexibility but also introducing a new set of security risks. Understanding these risks is crucial, as traditional organisational security measures often fail to address vulnerabilities specific to home and remote environments. This section provides context for the current gaps in the literature and highlights the need for a systematic review of threats targeting remote workers, together with the available mitigation strategies. While existing literature addresses organisational cybersecurity policies, there is limited focus on threats that specifically target remote workers (Mihailović et al, 2021). Empirical studies often overlook the risks associated with personal devices, home networks and informal working environments (Ferdousi, 2022). These shortcomings make it challenging to develop effective mitigation strategies that are tailored to the remote work setting. Mahboubi et al (2024) note that traditional threat models do not adequately account for vulnerabilities introduced by the remote working technologies, while Nwankpa and Datta (2023) emphasise that security awareness programs rarely address behavioural risks that target remote workers specifically. Zakki et al (2025) highlight that challenges such as technical, human and organisational nature do exist in the remote work environment, and these are challenges when it comes to delivering resource allocations to remote workers. Furthermore, the global survey by Gibson (2022) found that organisations that allow workers to bring personal devices to work often fail to implement comprehensive remote access policy protections. These studies reveal the lack of comprehensive reviews targeting remote workers addressing specific questions i) What are the main cybersecurity challenges that affect remote workers? and (ii) What mitigation strategies have been proposed?

This paper aims to bridge the identified gaps by conducting a systematic review of cybersecurity threats targeting remote workers. The review synthesises findings from various studies to provide a comprehensive understanding of the primary challenges and the effectiveness of proposed mitigation strategies. By addressing the research questions, this study contributes to both academic literature and practical applications in enhancing cybersecurity for remote work environments.

3. Research Methodology

The main objective of this paper was to tackle the urgent issues related to cybersecurity management in remote work settings. This systematic literature review followed PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses), a framework that ensures a rigorous and transparent methodology of synthesising research. The review process involved several steps, including identifying relevant studies, applying predefined inclusion and exclusion criteria, systematic data extraction and finally the synthesis of findings. Each step was conducted in accordance with PRISMA guidelines to ensure reproducibility and to minimise bias (Page et al., 2021).

3.1 Search Strategy

A comprehensive search of electronic databases was conducted, where the relevant literature was identified. The following databases were used: Google Scholar and Scopus. The search terms that were used were developed using the study's objectives and included the combination of keywords, "remote work AND cybersecurity", "hybrid AND cyber threats", "work from home AND security vulnerabilities", and "remote workers AND information security". Some Boolean operators, such as (AND, OR), were used to refine the search results and the retrieval of relevant studies. A total of 333 studies from Google Scholar and 4 from Scopus were returned in the initial search. After screening for the inclusion and exclusion criteria, a total of 20 studies were included in this systematic review.

3.2 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria applied in the selection review of the studies are outlined as follows: studies published between 2019 and 2025, including peer-reviewed works such as journal articles, conference proceedings, and peer-reviewed book chapters, must be in English and specifically address cybersecurity in the context of remote work. The exclusion criteria encompass opinions, blogs, and non-peer-reviewed materials, as well as studies that focus on general cybersecurity topics or are not available in full-text format. These criteria were assessed against a total of 337 papers prior to the final selection process.

3.3 Screening and Selection Process

The initial search yielded 337 articles, some of which were duplicates, some sources were not available in full text and only abstracts were available. After removing duplicates, 238 unique records remained. Screening abstracts reduced this to 118 studies for full-text review. Following full text assessment based on predefined inclusion criteria namely, relevance to the research objectives, empirical evaluation and alignment with the study scope, 20 studies were selected for this study.

3.4 Data Extraction

From each chosen source of the 20 papers, key information was systematically extracted and organised into a table that includes Authors and the year of publishing, the study's key findings, the cybersecurity threats addressed and the mitigation strategies.

3.5 Synthesis and Findings

The extracted data from the 20 papers were thematically analysed by the most common cybersecurity threats targeting remote workers, a common pattern among the studies and the gaps in literature. The findings were then organised into key themes, including social engineering and phishing, malware and ransomware, VPN and remote access, insider threat and human error. The mitigation strategies were summarised according to human and technical approaches. This methodology ensured a comprehensive and replicable review of the current literature on cybersecurity threats facing remote workers.

4. Findings

The systematic review of 20 peer-reviewed studies specifically addressing cybersecurity threats targeting remote workers and their mitigation strategies revealed several recurring threats. The studies differ in scope, methodology, and focus, but they highlighted the same threats, such as phishing, malware, ransomware, and insecure networks, which were the most frequently reported threats (Cuyugan and Rey, 2024; Mandadi et al, 2024; Pranggono & Arabo, 2021; Nwankpa et al, 2022). Human factors, such as employee behaviour, awareness, and attitudes, also play a significant role (Rakha, 2023; Qollakaj et al, 2025). Phishing attacks were reported in 15 studies, malware and ransomware in 12 studies, insecure networks and VPNs in 8 studies, and human factors such as the lack of security awareness in 7 studies. Table 1 below presents some of the major threat types together with the key findings of the studies.

Table 1: Major threats and mitigation strategies

Author (s)	Threat type	Mitigation strategy	Key findings
Mandadi et al (2024)	Spyware, Trojan horse, phishing, and insecure Wi-Fi	VPNs, strong passwords, endpoint management, training	Employee awareness and training is the most critical mitigation strategy
Cuyugan & Rey (2024)	Malware infections, Phishing	Policies, strong cybersecurity culture, and MFA deployment	Importance of multi-factor authentication and Zero Trust Network Access, MFA
Pranggono & Arabo (2021)	Phishing, Ransomware, DDoS, malware	Policies, employee education, VPNs, MFA, segmentation	Zero Trust approach is emphasised; remote work raises cybersecurity risks
Atstāja et al (2023), Georgiadou et al (2021)	Human factors, security awareness, phishing, and insecure networks	Cybersecurity training, CAT frameworks, and individual security focus	Human behaviour is key to information security; precaution-taking is critical
Nwankpa et al (2022); Milson & Altan (2023)	Confidentiality, data integrity, phishing, social	Cybersecurity training, MFA, VPNs, and patch management	Remote work increases risk; fostering cybersecurity best practices mitigates threats

Author (s)	Threat type	Mitigation strategy	Key findings
	engineering, and insecure home networks		
Hasan & Elweskiöld (2021); Qollakaj et al (2025)	Isolation, phishing, delayed support, invisible IT presence	Tailored phishing simulations, training and support	Remote work diminishes information flow; isolation increases phishing susceptibility
Cremer et al (2022); Alawida et al (2023); Rakha (2023)	Ransomware, malware, DDoS, phishing, hacking	Proactive cybersecurity, policies, guidelines, cyberattack detection & response	Proactive measures and legal compliance improve security in remote work
Bispham et al. (2021); Mahyoub et al. (2024)	Zoom bombing, ransomware, malware, phishing, Denial of service attacks (DDoS), physical device security	Policies, developing an incident response plan, VPNs, avoiding public WiFi, anti-Virus Software	The need for a zero-trust approach is emphasized, organisations should prioritize promoting security behaviour
Whitty, Moustafa and Marthie (2024); Nurse et al. (2021); Ozer et al. (2024)	Spear phishing, financial fraud, DDoS, insecure WiFi networks, VPNs, social engineering scams	Avoiding a one-size-fits-all approach, cybersecurity training, applying zero trust framework, regular updates on emerging threats	Remote work requires additional cybersecurity practices, the tension between workers wanting privacy and employer's goal to monitor workers, organizations to remain vigilant and adapt their cybersecurity strategies as the remote work trend increases
Khatri, Cherukuri and Kamalov (2023), Hiji and Alam (2022); Jerotirmay (2020)	Identity theft, DDoS, hacking, phishing, digital stalking, social engineering, lack of cybersecurity awareness, malware, data breaches, unsecure home networks	Regular cybersecurity awareness training, using VPNs, DNS security, data loss prevention, zero trust architecture	The study highlights the urgent need to improve cybersecurity hygiene practices and secure cyber infrastructure to mitigate future threats. Organisations must implement CAT framework to train and provide awareness to their employees regarding cyber threats. Traditional cybersecurity measures are no longer effective to protect organisations data.

As indicated in Table 1 above, several commonly recommended mitigation strategies encompass employee training and awareness programs, the use of Virtual Private Networks (VPNs) (Mandadi et al, 2024; Nwankpa et al, 2022; Milson & Altan, 2023), multi-factor authentication, the implementation of robust policies and guidelines (Pranggono & Arabo, 2021; Milson & Altan, 2023; Alawida et al, 2023), effective endpoint management, and the adoption of a zero-trust framework Atstāja et al, 2023; Georgiadou et al, 2021; Alawida et al, 2023). Numerous studies highlight the importance of employee education and proactive security measures, particularly in remote work environments (Alawida et al, 2023; Rakha, 2023). Specifically, 16 studies advocate for the utilisation of VPNs, while secure network configurations are endorsed by 10 studies. Additionally, 8 studies emphasise strategies such as multi-factor authentication and endpoint management, with 6 studies recommending the establishment of comprehensive policies and frameworks. Overall, the findings indicate that while technical solutions like VPNs and antivirus software play a crucial role, human-centric approaches and organisational policies are essential for achieving effective cybersecurity in remote work settings.

5. Discussion

The findings of this systematic review illuminate the urgent and multifaceted cybersecurity challenges encountered by remote workers, specifically highlighting ransomware, insecure networks, malware, and phishing as the most prominent threats in this realm. This observation aligns with previous research, including the comprehensive study conducted by Qollakaj et al. (2025), which reveals that the transition to remote work not only heightens the risk landscape but also exposes organisations to a diverse array of technical vulnerabilities and human-centric threats. A particularly critical aspect of this issue is the human factor, where deficiencies in

cybersecurity awareness among employees significantly compromise the safeguarding of digital assets. Research indicates that employees often lack a fundamental understanding of best practices in online security, making them potential targets for social engineering attacks. While implementing robust technical measures—such as leveraging Virtual Private Networks (VPNs) for secure data transmission, effective endpoint management to ensure devices are protected, and Multi-Factor Authentication (MFA) to add layers of security—are indeed essential for enhancing an organisation's cyber defences, literature consistently underscores the necessity of comprehensive training and awareness programs tailored for employees.

These educational initiatives have emerged as some of the most effective strategies to mitigate cybersecurity risks, emphasising the importance of cultivating a culture of security awareness within organisations. Successful cybersecurity frameworks, therefore, should seamlessly integrate both behavioural aspects, such as employee training and prompt reporting mechanisms for suspicious activities, alongside technological approaches to establish a formidable defence strategy. Moreover, this review highlights that organisations adopting a zero-trust security framework—which operates on the principle of never assuming trust, regardless of the user's location—coupled with well-defined policies and guidelines, are significantly better positioned to tackle the unique risks associated with remote work environments. By prioritising both sophisticated technical safeguards and ongoing employee education and engagement, companies can not only enhance their cybersecurity posture but also foster a resilient and adaptive cybersecurity culture that effectively addresses emerging threats in an increasingly digital landscape.

Numerous studies have identified significant challenges in the implementation of security measures for remote work, including resistance to change among employees, concerns surrounding privacy, and the varying nature of home office environments. These findings indicate a pressing need for flexible security policies that can be customised to address the diverse requirements of different remote work settings. While much of the existing research has concentrated on well-known threats such as malware and phishing, there is a noticeable gap in the literature regarding emerging threats, particularly social engineering attacks that target collaboration platforms and cloud-based applications. This oversight presents valuable opportunities for future research endeavours aimed at investigating advanced threat detection methodologies, the development of proactive employee monitoring systems, and organisational tactics designed to mitigate risks associated with remote work.

The ongoing discussion underscores the importance of a multifaceted approach that integrates policy frameworks, human-centred interventions, and robust technical safeguards as integral components for ensuring the security of remote work environments. Organisations must remain committed to evolving their cybersecurity strategies, not only to confront both emerging and existing threats but also to prioritise the continuous training and awareness of employees. This emphasis on ongoing education is crucial in fostering a culture of security awareness, ultimately reducing vulnerabilities in an increasingly digital and remote workforce.

6. Future Work

While the reviewed studies offer valuable insights into cybersecurity, several critical gaps persist that warrant further examination:

- Longitudinal Studies: A limited number of studies have explored the long-term effectiveness of various training programs and technical interventions. Understanding the sustained impact of these initiatives is essential for developing enduring cybersecurity strategies that can adapt to evolving threats over time.
- Focus on Small and Medium Enterprises (SMEs): The majority of existing research tends to concentrate on large organisations, leaving a notable deficit in studies that address the unique challenges faced by small to medium enterprises. SMEs often have distinct resource constraints and risk profiles, and targeted research in this area could lead to more effective and scalable cybersecurity solutions for these entities.
- Investigating Emerging Threats: The landscape of cybersecurity threats is continually shifting, particularly with the rise of new technologies such as collaboration platforms and the Internet of Things (IoT). However, there is a need for more comprehensive research into the vulnerabilities associated with these emerging technologies, as they pose unique risks that require tailored responses.

Addressing these gaps is crucial for organisations aiming to develop more agile and effective cybersecurity frameworks that cater specifically to the needs of remote workers. Furthermore, it will guide future research efforts in this essential field, ultimately enhancing overall cybersecurity resilience.

7. Conclusion

The systematic review was aimed at examining the main cybersecurity challenges facing remote workers and their mitigation strategies proposed in recent literature. The findings revealed that phishing, malware, ransomware and insecure networks remain the most persistent threats, while human factors such as the lack of cybersecurity awareness and training further amplify the risks in the remote work environments. Overall, by consolidating the findings across 20 studies, this paper contributes to the growing body of knowledge on remote work cybersecurity and identifies areas where research remains limited, particularly around emerging threats in collaboration tools and cloud platforms. Future research should investigate the advanced detection mechanisms, training models and long-term effectiveness of the zero-trust architecture in the remote work setting.

Ethics Declaration: Ethical clearance was not required for this research, as it was based solely on a systematic review of published academic and industry literature. No human participants or sensitive data were involved in the study.

AI Declarations: Grammarly was utilised to aid in the preparation of this paper. AI was engaged to assist with refining the language and editing the structure. The author maintained complete responsibility for reviewing, verifying, summarising, and interpreting all outputs, thereby ensuring the accuracy and integrity of the final document.

References

Admass, W.S., Munaye, Y.Y. and Diro, A.A. (2024). Cyber security: State of the art, Challenges and Future Directions. Cyber Security and Applications doi:https://doi.org/10.1016/j.csa.2023.100031.

Alawida, M., Omolara, A.E., Abiodun, O.I. and Al-Rajab, M. (2022). A Deeper Look into Cybersecurity Issues in the Wake of Covid-19: a Survey. Journal of King Saud University - Computer and Information Sciences, 34(10), pp.8176–8206. doi:https://doi.org/10.1016/j.jksuci.2022.08.003.

Aslan, Ö., Aktuğ, S.S., Okay, M.O., Yilmaz, A.A. and Akin, E. (2023). A Comprehensive Review of Cyber Security Vulnerabilities, Threats, Attacks, and Solutions. Electronics, 12(6), pp.1–42. doi:https://doi.org/10.3390/electronics12061333.

Atstāja, L., Rūtītis, D., Deruma, S. and Aksjoņenko, E. (2021). Cyber Security Risks and Challenges in Remote Work under the Covid-19 Pandemic. European Proceedings of Social and Behavioural Sciences, 121. doi:https://doi.org/10.15405/epsbs.2021.12.04.2.

Bispham, M., Creese, S., Dutton, W.H., Esteve-Gonzalez, P. and Goldsmith, M. (2021). Cybersecurity in Working from Home: an Exploratory Study. SSRN Electronic Journal. doi:https://doi.org/10.2139/ssrn.3897380.

Cremer, F., Sheehan, B., Fortmann, M., Kia, A.N., Mullins, M., Murphy, F. and Materne, S. (2022). Cyber risk and cybersecurity: A systematic review of data availability. The Geneva Papers on Risk and Insurance - Issues and Practice, 47(3), pp.698–736. doi:https://doi.org/10.1057/s41288-022-00266-6.

Cuyugan, M.J.R. and Rey, W.P. (2024). Beyond the Firewall: Strategies in Securing Remote Work Environment. 2024 14th International Conference on Software Technology and Engineering (ICSTE), pp.94–101. doi:https://doi.org/10.1109/icste63875.2024.00024.

Georgiadou, A., Mouzakitis, S. and Askounis, D. (2021). Working from Home during COVID-19 crisis: a Cyber Security Culture Assessment Survey. Security Journal, 35(35). doi:https://doi.org/10.1057/s41284-021-00286-2.

Gibson, K. (2022). Remote Work Creates a Gap in Security Practices - IT-Online. [online] IT-Online. Available at: https://it-online.co.za/2022/02/03/remote-work-creates-a-gap-in-security-practices [Accessed 29 Sep. 2025].

Hijji, M. and Alam, G. (2022) Cybersecurity Awareness and Training (CAT) Framework for Remote Working Employees. Sensors, 22, Article No. 8663. https://doi.org/10.3390/s22228663

Hasan, R., Elweskiöld, M. and Mutimukwe, C. (n.d.). Phishing in Remote Work: an Understanding of Factors That Enhance Employee Susceptibility to Phishing Attacks a Qualitative Study Degree Project at the Bachelor Level Computer and System Sciences Spring Term 2025. Available at: https://su.diva-portal.org/smash/get/diva2%3A1980971/FULLTEXT01.pdf? [Accessed 11 Sep. 2025].

Jerotirmay, J. (2020). View of Adapting to Remote Work: Emerging Cyber Risks and How to Safeguard Your Organization. Turcomat.org. Available at: https://turcomat.org/index.php/turkbilmat/article/view/15190/10904.

Khatri, S., Cherukuri, A.K. and Kamalov, F. (2023). Global Pandemics Influence on Cyber Security and Cyber Crimes. doi:https://doi.org/10.48550/arxiv.2302.12462.

Mahyoub, A., Luong, K., Aboutorab, H., Bui, H.T., Jarrad, G., Bahutair, M., Camtepe, S., Pogrebna, G., Ahmed, E., Barry, B. and Gately, H. (2024). Evolving Techniques in Cyber Threat hunting: a Systematic Review. Journal of Network and Computer Applications, 232, pp.104004–104004. doi:https://doi.org/10.1016/j.jnca.2024.104004.

Mandadi, S., Gochhayat, S.P., Torremocha, V. and Kethar, J. (2024). Cybersecurity Risks in Remote Work and Learning Environments and Methods of Combating Them. Journal of Student Research, 13(2). doi:https://doi.org/10.47611/jsrhs.v13i2.6808.

Mihailović, A., Cerović Smolović, J., Radević, I., Rašović, N. and Martinović, N. (2021). COVID-19 and Beyond: Employee Perceptions of the Efficiency of Teleworking and Its Cybersecurity Implications. Sustainability, 13(12), p.6750. doi:https://doi.org/10.3390/su13126750.

Milson, S. and Altan, B. (2023). Cybersecurity in Remote Work Environments: Challenges and Best Practices. easychair.org. Available at: https://easychair.org/publications/preprint/9lj5.

Newbold, J.W., Rudnicka, A., Cook, D., Cecchinato, M., Gould, S. and Cox, A.L. (2021). The new normals of work: a framework for understanding responses to disruptions created by new futures of work. Human–Computer Interaction, 37(6), pp.1–24. doi:https://doi.org/10.1080/07370024.2021.1982391.

Nurse, Jason R. C., Williams, Nikki, Collins, Emily, Panteli, Niki, Blythe, John and Koppelman, Ben (2021) Remote Working Pre- and Post-COVID-19: An Analysis of New Threats and Risks to Security and Privacy. In: 23rd International Conference on Human-Computer Interaction.

Nwankpa, J.K. and Datta, P. (2023). Remote Vigilance: the Roles of Cyber Awareness and Cybersecurity Policies among Remote Workers. Computers & Security, 130(1), pp.103266–103266. doi:https://doi.org/10.1016/j.cose.2023.103266.

Olson, M.H. (1983). Remote Office work: Changing Work Patterns in Space and Time. Communications of the ACM, 26(3), pp.182–187. doi:https://doi.org/10.1145/358061.358068.

Ozer, M., Kose, Y., Bastug, M., Kucukkaya, G. and Varlioglu, E.R. (2024). The Shifting Landscape of Cybersecurity: The Impact of Remote Work and COVID-19 on Data Breach Trends. arXiv (Cornell University). doi:https://doi.org/10.48550/arxiv.2402.06650.

Pranggono, B. and Arabo, A. (2020). COVID-19 Pandemic Cybersecurity Issues. Internet Technology Letters, 4(2). doi:https://doi.org/10.1002/itl2.247.

Qollakaj, K., Larsson, L.E. and Memeti, S. (2025). Cybersecurity of Remote Work migration: a Study on the VPN Security Landscape Post Covid-19 Outbreak. Array, 27, p.100437. doi:https://doi.org/10.1016/j.array.2025.100437.

Rakha, N.A. (2023). Ensuring Cyber-security in Remote Workforce: Legal Implications and International Best Practices. International Journal of Law and Policy, 1(3). doi:https://doi.org/10.59022/ijlp.43.

Sabin, J. (2021). The Future of Security in a remote-work Environment. Network Security, 2021(10), pp.15–17. doi:https://doi.org/10.1016/S1353-4858(21)00118-5.

Sanusi, M. (2025). Cybersecurity Challenges in Remote Work Environments: Common Threats and Mitigation Strategies. Available at: https://www.diva-portal.org/smash/get/diva2:1995670/FULLTEXT01.pdf [Accessed 28 Sep. 2025].

Schiliro, F. (2023). Towards a Contemporary Definition of Cybersecurity. Towards a Contemporary Definition of Cybersecurity. doi:https://doi.org/10.48550/arxiv.2302.02274.

Sokolic, D. (2022). Remote Work And Hybrid Work Organizations. [online] ResearchGate. Available at: https://www.researchgate.net/publication/359056200_Remote_Work_And_Hybrid_Work_Organizations.

Vivekananth, P. (2022). Cybersecurity Risks in Remote Working Environment and Strategies to Mitigate Them. International Journal of Engineering and Management Research, 12(1), pp.108–111. doi:https://doi.org/10.31033/ijemr.12.1.13.

Whitty, M.T., Moustafa, N. and Marthie Grobler (2024). Cybersecurity When Working from Home during COVID-19: considering the Human Factors. Journal of Cybersecurity, 10(1). doi:https://doi.org/10.1093/cybsec/tyae001.

Williams, J., Collins, N., Panteli, E., Blythe, N. and Koppelman, J. (2021). Remote Working Pre-and Post-COVID-19: An Analysis of New Threats and Risks to Security and Privacy. Communications in Computer and Information Science. doi:https://doi.org/10.1007/978-3-030-78645-8.

Zakki, M.N., Iftikhar, N., Khan, S.S.U., Nishat, F. and Arshi, O. (2025). Overcoming Challenges and Implementing Effective Information Security Policies for Remote Work Environments. Information Systems Engineering and Management, pp.135–151. doi:https://doi.org/10.1007/978-3-031-81481-5_10.

The Conceptualisation of a National Malware Intelligence Laboratory for South Africa

Wian Gertenbach, André McDonald, Mfundo Masango, Ele Mukondeleli, Ethan Buckinjohn, Rendani Mmbodi, Molebogeng Latakgomo, Ndabe Hhlongwane and Namosha Veerasamy
Council for Scientific and Industrial Research, Pretoria, South Africa

WGertenbach@csir.co.za
AMcdonald@csir.co.za
MMasango1@csir.co.za

Abstract: South Africa has become an increasingly attractive target for cybercriminals, with malware and ransomware attacks on critical national infrastructure and public institutions rising in both frequency and severity. High-profile incidents, such as ransomware attacks on Transnet, the Department of Justice, and the Government Employees Pension Fund (GEPF) highlight the scale of the threat. These attacks result in severe operational disruption and financial losses, often due to inadequate readiness and response mechanisms. This paper presents the conceptualisation of a national malware intelligence laboratory (NMIL) for South Africa, designed to strengthen domestic readiness and response to malware-related threats. A literature study was carried out to determine the main gaps. This was complemented by a stakeholder interview and questionnaire. This led to a gap analysis, and the examination of existing models as reference for the proposed design of the NMIL. This process not only identifies systemic weaknesses in the national cyber defence capabilities of South Africa but also evaluates how an NMIL could be integrated into existing national cybersecurity processes. The gap analysis revealed limited coordination of malware intelligence sharing across sectoral computer security incident response teams (CSIRTs), the potential to improve the ability of the National Cybersecurity Hub (CSHUB) to aggregate and disseminate actionable threat data, and insufficient hands-on exposure to malware in current cybersecurity education and training programs. In response, the paper introduces a framework and reference model that defines NMIL functions and its manner of integration into the national cybersecurity ecosystem. Specifically, the laboratory would provide high-quality malware intelligence to the CSHUB, including sample analysis results, threat profiling, and advisory support on removal tools, to improve effective response coordination. Additionally, the laboratory would offer access to a sandboxed training environment to educational institutions, thereby adding greater depth to cybersecurity education and promoting national cybersecurity readiness. The framework and reference model is developed using a systems engineering approach, to detail the NMIL's information flows, interfaces and functional domains. It is anticipated that the formal process resulting in conceptual laboratory provides a replicable approach for institutionalising national malware laboratories. This model offers both strategic and operational insights for South Africa as well as other developing countries.

Keywords: Malware analysis, Malware threat intelligence, Malware laboratory, Skills development

1. Introduction

In the last decade, malware attacks have evolved significantly in sophistication, resulting in a global influx of attacks on critical infrastructure and services. Emerging trends in the adoption of sophisticated criminal business models – particularly those leveraging affiliate distribution networks that function as a malware-as-a-service ecosystem – pose a significant threat to developing countries that lack adequate defence capabilities (Gajjar and Taherdoost, 2024). This threat is realised in South Africa, which faces an annual increase in malware attacks as criminals shift their focus more towards developing countries (Henrico and Els, 2025). A national survey conducted by the Council for Scientific Research (CSIR) in 2024 found that malware attacks were the most common incident type reported by South African organisations, accounting for 65% of incidents reported (Ngejane, 2024). Ransomware is especially prevalent, with attacks on South African public entities such as Transnet, the Department of Justice, and the Government Employees Pension Fund (GEPF) documented and publicised in the last five years (Greig, 2024; Moyo, 2024; Timcke et al., 2023). In a survey conducted by Sophos in 2025 on 154 South African organisations that were hit by ransomware in the previous year, 58% of respondents stated that the most common cause of attacks was a lack of expertise (Sophos, 2025). Interpol reported that ransomware remains one of the most prevalent forms of cybercrime across Africa, with South Africa ranking second in the number of ransomware threats detected in 2024 (Interpol, 2025).

Despite legislative advancements in South Africa such as the establishment of the Cybercrimes Act (Act 19 of 2020), and National Cybersecurity Hub (CSHUB), systemic gaps remain in malware threat intelligence sharing, malware analysis capacity, and the integration of hands-on training (Henrico and Els, 2025). These challenges signify the need for improved national malware threat intelligence, stronger national coordination, and investing in hands-on cybersecurity training to develop skills in digital forensics, malware analysis and open-source

intelligence (Interpol, 2025). To address these gaps, the paper introduces the conceptualisation of a South African national malware intelligence laboratory (NMIL) as a critical component in advancing domestic readiness against malware-related threats and a supportive capability that strengthens upstream defensive functions. A structured process was applied, using a cause-effect analysis informed by literature, inputs from academic and cybersecurity professionals through interviews and questionnaires, to identify the root causes of poor malware readiness. This led to the conceptualisation of a laboratory model to address these root causes.

The paper presents several contributions. The process through which the NMIL was conceptualised, as presented in section 2, is replicable and can be used in different scenarios. The gap analysis, presented in section 3, identifies specific gaps that arise within the South African cyber defence context and was used to conceptualise the proposed solution. In section 4, a framework that defines a solution (the NMIL) to address the gaps identified, as well as a reference model that serves as a blueprint for implementing and validating NMILs in general, is developed. Section 5 explains the limitations of the study as well as recommendations and future work.

2. Process for Conceptualisation of a National Malware Intelligence Laboratory

The conceptualisation of the NMIL requires a structured process to ensure that the proposed NMIL framework is evidence-based and addresses critical gaps that are contextually relevant in South Africa. The process combines theoretical perspectives on the state of domestic malware-related attacks, practical examples from international laboratory models that demonstrate clear benefits, and an examination of domestic realities through an interview and questionnaire. This approach ensures that the NMIL framework reflects international best practices while simultaneously remaining responsive and relevant to local challenges and constraints.

The first step involved a literature survey to understand the local need for addressing the increase in malware-related attacks and to identify critical gaps in domestic cybersecurity capabilities. Several sources were studied to ensure maximal coverage. Industry reports from cybersecurity vendors, academic publications, news articles, and reports by research organisations provided practical insight into the South African threat landscape and skills shortages. International technical architectures and implementation approaches for cybersecurity laboratories and malware analysis capabilities were studied to understand how other nations structure their malware intelligence capabilities and integrate them with existing cybersecurity frameworks. The review served to identify the need for a practical solution, like a national malware laboratory, to support malware intelligence sharing and analysis in national resilience, and how its functions may integrate into existing national computer security incident response teams (CSIRTs) and related institutions. This step provides the conceptual foundation to identify the functions that an NMIL should perform, specifically in the domains of malware threat intelligence sharing and addressing critical skills shortages. The outputs of the literature review were synthesised into a cause-effect diagram to visualise relationships between the root cause (malware-related incidents in the public sector) and underlying systemic gaps.

The cause-effect analysis led to the identification of two critical gaps that contribute to the country's struggling cyber resilience, namely the lack of relevant skills to address malware-related attacks, and ineffective coordination of malware threat intelligence sharing between sectoral CSIRTs. These gaps were explored in a gap analysis to fully understand their effects on national resilience. The gap analysis was supported by a questionnaire and interview with stakeholders across research institutions and academia. The goal was to obtain additional evidence – not publicly available form academic publications or articles – from industry experts who work closely with entities that play a role in threat intelligence sharing, as well as to understand the role of malware-related content in formal education. Although a broader set of stakeholder engagements were planned, only one in-depth interview and one academic questionnaire response were successfully completed during this phase. As the study was explorative and conceptual in nature, these engagements were treated as validation inputs rather than statistically representative samples.

The interview provided qualitative insight into operational challenges faced by national cybersecurity teams, specifically coordination between CSIRTs in the public sector and the lack of available malware analysis infrastructure. The academic response was used to confirm challenges related to malware content in higher education curriculums, specifically the lack of skilled lecturers and adequate infrastructure for practical training.

The structured process established a solid foundation to derive a suitable framework and reference model for a NMIL. The framework defines the functions of the NMIL that address the gaps identified. This includes the NMIL's core domains of operation to establish it as a technical, strategic, and educational hub. The reference model provides a system representation of the NMIL's inputs, processes, outputs, and information flows. The

framework together with the reference model serves as a blueprint for how the laboratory would interface with its stakeholders and integrate into South Africa's cybersecurity ecosystem.

3. Findings: Gap Analysis and Root Causes

The rise in successful malware-related breaches in the South African public sector can be attributed to various causes. During the gap analysis, various causes were identified and investigated. The most significant cause is the public sector's poor malware readiness posture (Siphambili et al., 2024). Evidence for this cause is the number of successful malware-related attacks on critical public entities in recent years. A few high-profile attacks that have occurred in recent years includes the City of Johannesburg, the Department of Justice, GEPF and Transnet (Moyo, 2024; Ngejane, 2024; Pieterse, 2021; Siphambili et al., 2024; Timcke et al., 2023). The reasons for these attacks vary per incident. Henrico and Els (2025) note that the main cause stems from systemic weaknesses and a lack of relevant skills to protect and respond to incidents. Section 3.1 explores the systemic weaknesses, evaluating the challenges faced by current systems and entities to combat malware-related attacks. Section 3.2 explores the reason why skills shortages in the malware analysis domain are a reality.

3.1 Intelligence Sharing and Infrastructure Gaps

An important question to consider is whether systems, processes and policies are already in place to prevent the rise of successful malware-related attacks, and whether they are adequate. As a central role player, the South African National CSHUB acts as an advisory to other sectoral CSIRTs to improve the effectiveness of incident response activities. The CSHUB is mandated under the National Cybersecurity Policy Framework (NCPF) to focus on collaboration with stakeholders in the public and private sectors to facilitate incident coordination, information dissemination, public cybersecurity awareness and the promotion of national standards (State Security Agency, 2015). Unfortunately, various challenges exist in this coordination plan. To uncover the specific reasons, the authors interviewed a key stakeholder of the CSHUB to provide insight into the current challenges it faces.

The most pressing challenge currently faced by the CSHUB and other CSIRTs is the effective distribution of local malware intelligence, according to (Ngobeni, 2025), the interviewee. Although infrastructure and standard operating procedures (SOPs) for threat intelligence sharing exist, their adoption is still in an immature phase. There is an overall reluctance from public and private institutions to participate in intelligence sharing, mainly due to the fear of reputational damage and competitiveness. As a result, current infrastructure and processes that facilitate intelligence sharing are not fully utilised, thereby leading to the neglect of sustained operations. Furthermore, there is insufficient integration of threat intelligence sources with the CSHUB, which hinders their ability to share critical information that may help to prevent malware-related incidents. Specific sources include findings from dark web monitoring, which is rich in unconventionally obtainable information and resources. This integration could greatly strengthen the CSHUB's position as a hub of proactive intelligence.

Ngobeni (2025) explains that South Africa's cybersecurity infrastructure is still at an infant stage, and official government support structures are inadequate to aid or resolve critical issues. This leads the departments to rely on weak internal IT support (this also points to a weak cybersecurity workforce) rather than robust incident response units. In incidents involving malware, other service providers are rather employed. However, their services are typically too costly for government departments, and security clearance requirements can also pose challenges. Additionally, there is a noted absence of adequate virtual infrastructure for testing and compliance, indicating current setups are insufficient for proactive malware handling or simulation. Overall, these issues are still neglected due to the NCPF being a non-binding policy rather than law, limiting enforcement and development of the necessary infrastructure and processes.

3.2 Skills and Educational Gaps

A critical factor contributing to South Africa's malware-readiness challenges is its increasing cybersecurity skills gap (de Jager et al., 2023). There is a significant lack of technical skills in cyber forensics and threat intelligence, especially under recently graduated individuals entering the workforce. Thomson et al. (2019) explain a few causes, including the insufficient integration of general cybersecurity education in undergraduate computer science curricula, a lack of educators specialised in cybersecurity, and a curriculum overload, which prevents the addition of more content. Practical, hands-on training is also not featured in most curriculums, as undergraduate programs focus mostly on theoretical concepts (de Jager et al., 2023). There is a shift in modern education to include more experimental and practical training for self-learning, which is currently lacking in tertiary education for cybersecurity-related topics (Steyn et al., 2021). A questionnaire was provided to experts in the educational field with the purpose of gaining additional input, (apart from the literature study), on the current level of

exposure that computer science and information technology students receive regarding cybersecurity, and especially malware concepts. The questions were designed to explore how academic programs address malware-related topics, how well students are prepared for real-world incidents, and what kind of institutional support could improve teaching and research in this domain.

The first question aimed to determine at what stage of tertiary education students are introduced to malware-related content, such as detection, analysis, and reverse engineering. The respondent indicated that such exposure typically occurs only at the Honours level and is limited to those who specifically choose a cybersecurity specialisation. This highlights that undergraduate programs lack the incorporation of cybersecurity topics, including malware, leaving most students without foundational knowledge in this area.

The second question assessed how effectively current curricula prepare students to deal with real-world cyber threats like ransomware and malware-as-a-service. The respondents noted that students are not sufficiently exposed to these advanced and evolving threat models. One of the main reasons cited was the difficulty in balancing a wide range of required topics within the Association for Computer Machinery (ACM)-aligned CS curriculum. This is strengthened by Thomson et al. (2019)'s findings on curriculum overload, and suggests that while institutions follow recognised guidelines, practical threat engagement is often exchanged for theoretical or foundational topics.

The third question focused on identifying the main skill gaps among graduates entering the cybersecurity workforce. The respondent revealed that the most pressing issue is the lack of real-world, hands-on experience. The current education model's focus is on breadth (broad, theoretical knowledge) over depth (practical, hands-on skills), leaving graduates unprepared for workplace demands (Manson and Pike, 2014). Graduates may understand theoretical concepts but often struggle to apply them in practical malware analysis or defence contexts, underscoring the need for experiential learning opportunities.

The fourth question explored the perceived value of integrating malware-related content into computer science curricula and asked for suggestions on how to effectively support this inclusion. The respondent mostly supported the idea, with one recommending the introduction of general malware knowledge at the undergraduate level and more advanced, practical work at the postgraduate level. There was an emphasis on the need for well-equipped laboratory environments where students can safely conduct malware analysis. The respondent also pointed to a shortage of qualified lecturers in this field, highlighting a broader capacity issue within academia. This challenge is highlighted by Thomson et al. (2019), who note that some computer science educators are not specialised in security, and may not have the necessary knowledge to effectively integrate and teach cybersecurity topics.

The final question inquired whether access to a national malware intelligence laboratory, with features like anonymised datasets, sandbox environments, expert guidance, and training resource, would add value to current curricula. The respondent agreed that such an initiative would directly address the current lack of suitable laboratory environments mentioned earlier. This indicates strong support for collaborative, centralised infrastructure that could enhance both education and research capabilities in malware analysis.

The questionnaire response and literature findings reveal a consistent theme: while there is academic interest in malware-related topics, exposure remains minimal, practical experience is limited, and institutional capacity is constrained. A national malware intelligence laboratory could therefore serve as a crucial bridge between theory and practice, equipping students and educators with the necessary tools, data, and expertise to become a capable workforce in South Africa.

3.3 Summary of Gaps

The gap analysis showed that various factors contribute to the country's struggle to combat malware-related attacks. Cybersecurity infrastructure and intelligence sharing processes are not optimised or utilised to its full potential, which ultimately results in private and public sector entities seeking solutions from external service providers. For government departments, this is a major challenge, as reliance on private service providers is often constrained by limited budgets, the confidential nature of projects, and the strict clearance requirements involved. The lack of specialised cybersecurity skills, particularly in incident response and malware analysis, leads to a decrease in readiness. Recently graduated cybersecurity practitioners often lack the practical knowledge needed to be applied to complex situations, which stems from the limited exposure to practical scenarios involving malware during their studies. In section 4 a solution is proposed that directly addresses the identified gaps in a framework that can be replicated by countries facing similar issues.

4. Proposed Framework: National Malware Intelligence Laboratory (NMIL)

The NMIL is conceptualised to address the identified gaps outlined in section 3. The gaps are classified in three groups, namely infrastructure and tooling gaps, intelligence sharing gaps, and skills and educational gaps. Table 1 shows a mapping of gaps to the direct NMIL response and its outcome. This structure helped guide the derivation of the NMIL's functional domains, which is based on a similar design described by Svajcer (2015). Although not addressing the same gaps as the NMIL, Svajcer (2015) explains the typical requirements for a malware laboratory, with emphasis on malware sample acquisition from diverse sources, analysis systems, and end-user systems. The NMIL follows a similar structure but includes an important intelligence generation function and educational system.

Table 1: Identified Gaps Mapped to the NMIL's Response and Expected Outcomes

Identified Gap	NMIL Response	Expected Outcome
Infrastructure and tooling gaps.		
Absence of virtualised infrastructure for tool testing and compliance.	Dedicated virtual environment where malware removal tools can be evaluated and developed to comply with national standards.	Capability to safely test and validate malware removal tools, ensuring readiness for real-world deployment.
Lack of hands-on malware analysis in prevention and response to public incidents.	Dedicated malware collection capability and sandbox environments for analysis.	Improved national capacity for malware research, enabling effective knowledge sharing.
Intelligence sharing gaps		
Lack of rich information sources.	Dark web monitoring and information gathering from various sources.	Broader and more comprehensive threat intelligence coverage, resulting in early threat detection.
Lack of specialised hardware to facilitate intelligence sharing.	Platform to process intelligence from a variety of sources and share it with the CSHUB to distribute to sectoral CSIRTs.	A robust intelligence processing system that enables efficient data correlation and dissemination among stakeholders.
Skills and educational gaps.		
Lack of practical training in universities.	Development of practical training material, in the form of laboratory scenarios.	Increased number of graduates with practical malware analysis experience.
Lack of access to infrastructure for practical training.	Provides universities with access to virtualised infrastructure to experiment and expose students to malware.	Improved access to specialised environments for students, lecturers and researchers to enhance skill development.

Existing malware laboratories—such as those operated by ANSSI (France), JPCERT/CC (Japan) and the United States Cybersecurity and Infrastructure Security Agency (CISA)—focus on technical analysis and indicators of compromise (IOC) dissemination within their national ecosystems. Academic labs address training or research functions separately. However, no existing models integrate malware intelligence, national coordination, and practical skills development in a manner tailored to South Africa's unique challenges. The NMIL therefore adapts international laboratory principles to address these specific national gaps.

In short, the NMIL would function as a technical hub for malware analysis and intelligence production as well as an educational hub for practical training and capacity building. Section 4.1 details the architecture of the reference model, which translates its functional domains into a system-level representation. Its integration with external systems is explained in section 4.2 , with information flows between systems clearly defined.

4.1 Conceptual Architecture

The NMIL operates in four key functional domains, namely malware collection, preservation and analysis, threat intelligence generation and sharing, education and training, and tool development and innovation. Figure 1

shows a diagram of the internal NMIL system and how the different functional domains interact with each other. Each functional domain consists of different sub-system modules that realise the goals of the NMIL.

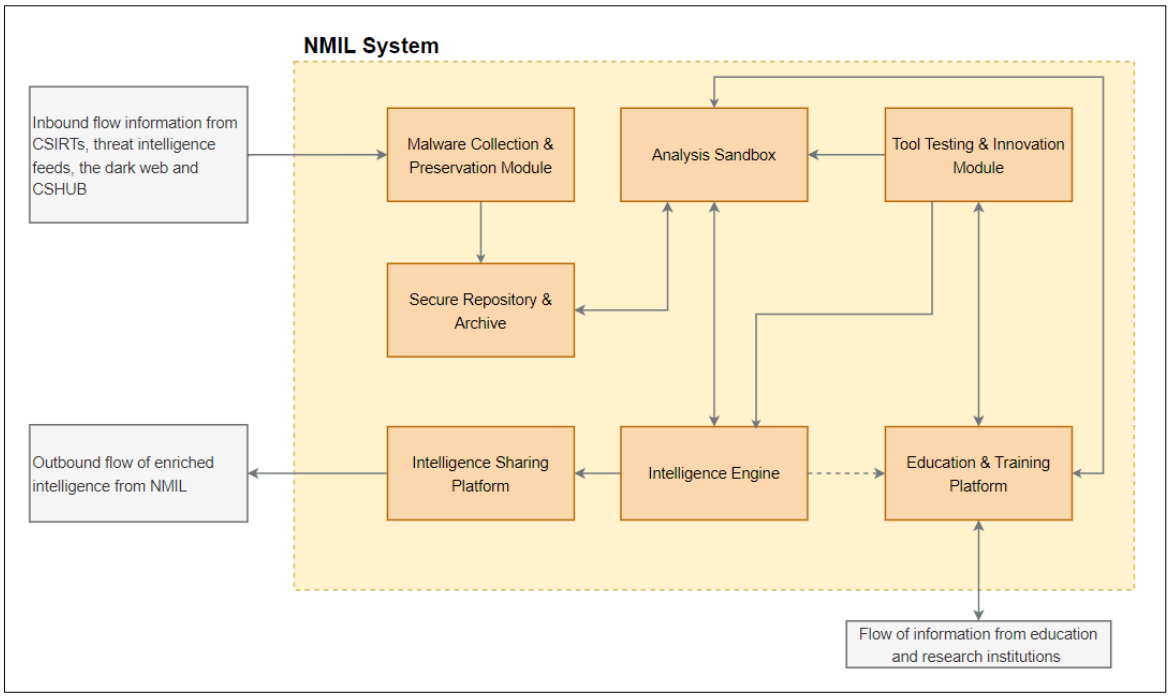


Figure 1: Internal NMIL System with Functional Domains

4.1.1 Malware collection, preservation and analysis

The NMIL is equipped with tools and hardware to collect, preserve and analyse malware, which ensures the acquisition and curation of malware samples from diverse sources, including the dark web, incident reports, the CSHUB, and international threat feeds. Collected samples are maintained in a secure repository with metadata describing their origin, behaviour, and IOCs. Analysis of these samples involve applying static and dynamic analysis techniques to dissect malware and understand its execution patterns, persistence mechanisms, and attack vectors. Through this process the NMIL generates technical insights in the form of analysis reports that enable the profiling malware to be used as actionable intelligence.

The Malware Collection and Preservation Module's main responsibility is to gather, register, label, and prepare inbound malware samples for secure storage and future use. Each incoming sample is assigned a unique identifier and labelled with essential metadata, including its type, hash value, origin, timestamp and source. This process ensures sample traceability from the point of origin and establishes a digital record of provenance and chain of custody for forensics and analysis purposes. A typical requirement for malware laboratories is the ability to collect malware as soon as it is located publicly or on the dark web (Svajcer, 2015). This module interfaces with the dark web and constantly monitors and scrapes information and malware samples to be processed by the NMIL. Registered samples are transmitted to the Secure Repository and Archive to maintain a catalogue of all collected samples. Transmission takes place over a secure data exchange channel to ensure that all content is isolated from other segments of the NMIL network. This prevents potentially dangerous artifacts from compromising the integrity of the NMIL. The repository enforces strict access and transfer protocols to prevent unauthorized handling or accidental exposure.

The repository interfaces directly with the Analysis Sandbox. The sandbox is an isolated, virtual environment designed to safely execute and observe malware behaviour without risking contamination of other sub-systems. It allows the simulation of different reproducible virtual environments to monitor malware behaviour in specific scenarios. This is ideal for testing new and existing malware removal and defence tools, which are developed and documented in the Tool Testing and Innovation Module. The sandbox is equipped with the necessary tools to perform static, dynamic, and code analysis. Analysts and researchers, with the required access rights, can copy samples from the repository to the sandbox to run experiments, test tools, perform research and return structured behavioural reports back to the repository once analysis is complete. Analysis results and reports also

feed into the Intelligence Engine, where observed tactics, techniques and procedures (TTPs) are mapped against frameworks such as MITRE ATT&CK (Svajcer, 2015).

The sandbox also interfaces with the Education and Training Platform. This allows access for registered students and instructors to experiment with malware practically without the risk of compromising their personal computers or equipment. The sandbox's function is thus twofold: To provide an operational environment to generate technical results for critical stakeholders, and to provide an experimental and learning environment where researchers and students can freely test, learn, and teach the principle of malware analysis.

4.1.2 Threat intelligence generation and sharing

This domain is responsible for the generation and sharing of actionable intelligence. The Intelligence Engine is responsible for the translation of technical findings from the Analysis Sandbox – such as behavioural indicators, IOCs and TTPs – into actionable insights by applying correlation, enrichment, and classification techniques to derive meaningful insights. The main goal of this module is to ensure that intelligence is operationally useful to security practitioners and NMIL stakeholders. Once validated, the refined intelligence is shared through the Intelligence Sharing Platform, ensuring secure transmission to external entities.

The Tool Testing and Innovation Module interacts with the Intelligence Engine by providing performance data and detection insights from various security tools – existing and custom developed – tested within the NMIL's environment. The Intelligence Engine processes these findings into actionable information, which can be shared to stakeholders in the form of tool recommendations.

Specific results processed by the Intelligence Engine are shared with the Education and Training Platform to be integrated with training material. In Figure 1, this interface is denoted with a dotted line and is one-directional. This means that the Education and Training Platform cannot access the Intelligence Engine directly. The Intelligence Engine shares information on a case-to-case basis, as some intelligence might be sensitive. As such, only specifically curated information is shared with the Training and Education Platform.

4.1.3 Support for education and training

The NMIL facilitates practical, hands-on knowledge transfer and training through its Education and Training Platform. It offers universities and academic institutions access to the Analysis Sandbox to support practical skills capacity building. The platform hosts various educational material developed from technical analysis findings, intelligence generation (carefully curated not to reveal sensitive stakeholder information), and tool experimentation, which is made available to educational institutions through university lectures and as online resources.

Additionally, research outputs are generated by the NMIL and published, facilitating academic collaboration. Leveraging real-world malware data and samples, the NMIL becomes a desired platform for experimental learning and a capable workforce development.

4.1.4 Tool experimentation, development and innovation

The sandboxed nature of the NMIL allows for experimentation with various malware removal and endpoint detection and response tools. Insights gained from testing these tools can be presented to stakeholders in the form of consultations. The Tool Testing and Innovation Module facilitates this capability and allows the research and development of new and existing malware prevention, removal and defence tools. It interfaces with the sandbox to allow the deployment of these tools in different scenarios. Documentation and tool-readiness reports feed into the Intelligence Engine, providing strategic insights to external entities on tool usage and deployment. The Tool Testing and Innovation Module also interfaces with the Education and Training Platform to allow students to experiment and learn how to use and develop industry-grade tools.

4.2 Data and Intelligence Flows

The diagram in Figure 2 illustrates the interaction and information flow between the NMIL and its key stakeholders in the South African cybersecurity landscape. At the centre, the NMIL System acts as the hub for malware intelligence generation and knowledge sharing. It interfaces with several external systems – its stakeholders – and facilitates the flow of information to and from the different systems.

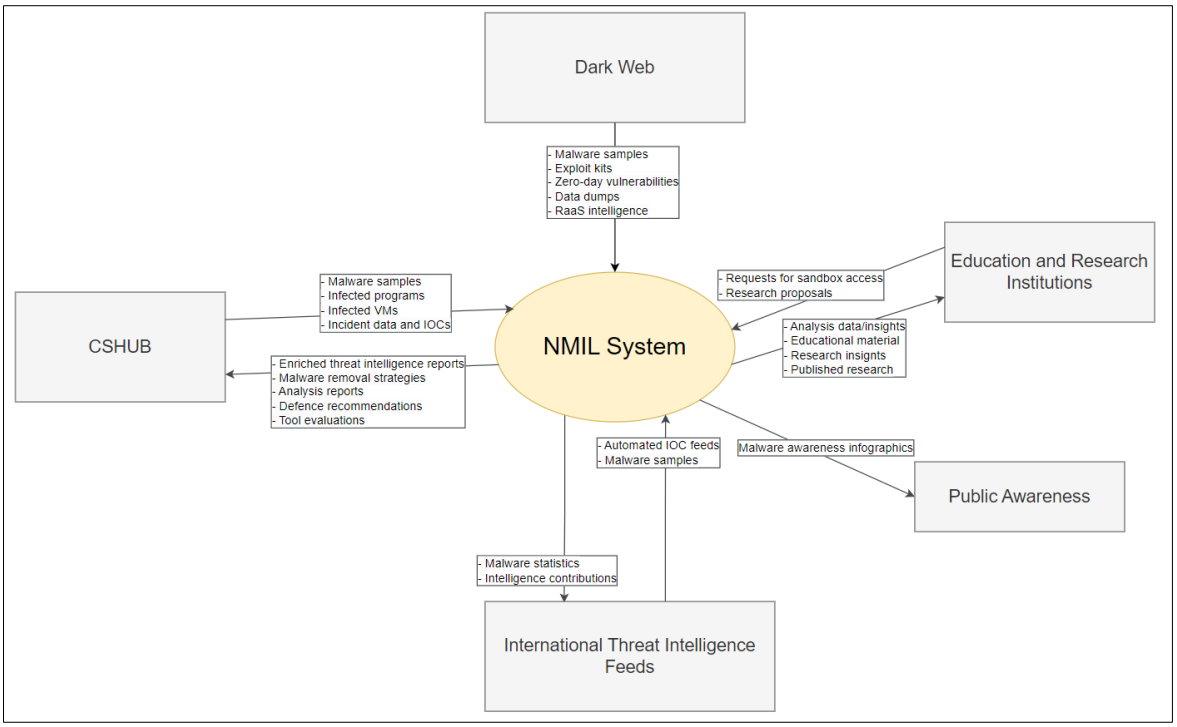


Figure 2: National Malware Intelligence Laboratory System Context Diagram

While the NMIL is not designed to function as a Security Operations Centre (SOC), its outputs naturally align with the higher-tier analytical functions commonly found in SOC structures. In a tiered SOC model, Tier 1 and Tier 2 teams typically perform alert monitoring, initial triage, and basic investigation (Cassetto, 2025). More advanced malware reverse engineering, deep-dive incident reconstruction, and threat intelligence correlation activities are performed by Tier 3 teams. In this context, the NMIL should be viewed as a national-level extension of Tier 3 analytical capability, rather than an operational SOC service. When organisations or sectoral CSIRTs escalate complex malware-related incidents that exceed their internal capacity, NMIL provides specialised analysis, tooling insights, and enriched intelligence. This positioning ensures the NMIL complements—rather than replaces—existing SOC and CSIRT functions, by supplying advanced technical outputs that feed back into frontline operational teams and national coordination structures such as the CSHUB.

The NMIL's interface with the CSHUB is bidirectional. It receives malware samples, infected programs and virtual machines (VMs), and incident data, which is then processed internally through its modules and platforms (see Figure 1) to produce actionable outputs. These elements can also be obtained from other sources, such as international threat intelligence feeds or dark web intelligence. The NMIL provides the CSHUB with enriched malware intelligence, removal and defence strategies, analysis reports, and tool evaluation results. This can then be used by the CSHUB to share with its own stakeholders (the sectoral CSIRTs), to fulfil its role in the national cybersecurity ecosystem.

The Dark Web acts as a source of threat intelligence, providing the NMIL with critical inputs which include malware samples, exploit kits, zero-day vulnerabilities, data dumps, and general information on the developments of ransomware-as-a-service. This information feeds into the NMIL for analysis and enrichment. Enriched intelligence is distributed to other systems such as the CSHUB and international threat intelligence feeds.

Education and Research Institutions interact with NMIL through requests for sandbox access, learning material and research proposals, enabling academic collaboration and experimental research on malware behaviour. In turn, NMIL shares analysis data and insights, learning material, research findings, and published papers, fostering a knowledge-sharing ecosystem that supports cybersecurity education and innovation. Knowledge-sharing is also enabled by the NMIL's interaction with the public. The NMIL produces malware awareness infographics for the public based on current threat trends and intelligence findings from international feeds and the dark web. This helps inform the public and organisations about emerging threats and best practices for protection.

The International Threat Intelligence Feeds component represents NMIL's external integration with global cybersecurity networks. NMIL provides malware statistics and intelligence contributions to international partners, while receiving automated IOC feeds and malware samples in return. This ensures South Africa's participation in the broader global cyber intelligence community, promoting mutual threat detection and prevention.

5. Limitations and Future Work

The findings presented in this paper are exploratory and primarily conceptual. Stakeholder input was based on one practitioner interview and one academic response, which limits generalisability of the results. However, these engagements provided valuable first-hand confirmation of the infrastructure, educational and information sharing gaps identified in the literature. The proposed NMIL framework should therefore be viewed as a proof-of-concept model grounded in initial evidence.

Future work will focus on system development and pilot integration with the CSHUB to assess its operational feasibility. Performance metrics will be defined to evaluate the NMIL's effectiveness. Further stakeholder engagement, particularly with educational institutions and the CSHUB, will support the NMIL's alignment with the national cybersecurity objectives and its integration into higher education for sustainable skills development.

6. Conclusion

This paper conceptualised a national malware intelligence laboratory (NMIL) for South Africa to address growing malware-related threats. A structured process was used by combining a literature review, a cause-effect analysis and limited stakeholder interaction to identify two key gaps in the South African cybersecurity landscape: insufficient malware analysis infrastructure and weak intelligence sharing, alongside limited exposure to practical training in computer science curricula.

The proposed NMIL integrates four functional domains: malware collection, preservation and analysis; threat intelligence generation and sharing; education and training; tool development and validation. These functional domains address the identified gaps through enhanced intelligence sharing mechanisms with the CSHUB and by developing a capable workforce through access to sandboxed training environments.

As a conceptual framework and model, the NMIL provides a replicable blueprint for enhancing cyber defence capabilities in other developing countries. Future work will focus on pilot implementation, expanded stakeholder engagement, and performance evaluation metrics to validate its operational and educational effectiveness.

Ethics Declaration: No ethical clearance was required for the completion of research activities mentioned in this paper.

AI Usage Declaration: Generative artificial intelligence (AI), specifically large language models (LLMs), were used to enhance writing, fix grammatical issues and provide summaries of literature for dissemination of information. No tools were used to write original content. AI tools used to aid the writing process include ChatGPT and Claude.

References

Cassetto, O., 2025. SOC Analyst Tier 1 vs. Tier 2 vs. Tier 3: Key Differences & Responsibilities [WWW Document]. Radiant Security.

de Jager, M., Futcher, L., Thomson, K.-L., 2023. An Investigation into the Cybersecurity Skills Gap in South Africa. In: Human Aspects of Information Security and Assurance, IFIP Advances in Information and Communication Technology. Springer, Cham, pp. 237–248.

Gajjar, V.R., Taherdoost, H., 2024. Cybercrime on a Global Scale: Trends, Policies, and Cybersecurity Strategies. In: Proceedings - 2024 5th International Conference on Mobile Computing and Sustainable Informatics, ICMCSI 2024. Institute of Electrical and Electronics Engineers Inc., pp. 668–676.

Greig, J., 2024. LockBit takes credit for February shutdown of South African pension fund [WWW Document]. The Record.

Henrico, S., Els, S., 2025. Cyber Attacks in South Africa: Geopolitical and legal implications. African Security Review 1–25.

Interpol, 2025. Interpol Africa Cyberthreat Assessment Report 2025.

Manson, D., Pike, R., 2014. The case for depth in cybersecurity education. ACM Inroads 5, 47–52.

Moyo, A., 2024. Justice department suffers another cyber attack [WWW Document]. ITWeb. URL https://www.itweb.co.za/article/justice-department-suffers-another-cyber-attack/rW1xLv5nJkx7Rk6m (accessed 9.15.25).

Ngejane, H., 2024. DATA BREACHES IN SOUTH AFRICA: SURVEY REPORT.

Ngobeni, S., 2025. Interview by W. Gertenbach. Unpublished interview conducted in Pretoria on 25 August 2025.

Pieterse, H., 2021. The Cyber Threat Landscape in South Africa: A 10-Year Review. The African Journal of Information and Communication 28, 1–21.

Siphambili, N., Mahlasela, O., Baloyi, E., Mukondeleli, E., 2024. A Review of the South African Public Sector's Capability in Combating Ransomware. In: 2024 4th International Multidisciplinary Information Technology and Engineering Conference (IMITEC). IEEE, pp. 493–499.

Sophos, 2025. The State of Ransomware in South Africa 2025.

State Security Agency, 2015. The National Cybersecurity Policy Framework.

Steyn, A.A., Botha, A.J.M., Coetzee, D., Villiers, M. de, 2021. Interactive Learning: Introducing a First-Year Systems' Analysis and Design Course. pp. 171–186.

Svajcer, V., 2015. Building a Malware Lab in the Age of Big Data. In: 25th Virus Bulletin International Conference. Prague.

Thomson, K.-L., Futcher, L.A., Gomana, L., 2019. Towards a framework for the integration of information security into undergraduate computing curricula. South African Journal of Higher Education 33.

Timcke, S., Gaffley, M., Rens, A., 2023. The centrality of cybersecurity to socioeconomic development policy: A case study of cyber-vulnerability at South Africa's Transnet. The African Journal of Information and Communication (AJIC) 1–28

Cyber–Space Governance: Securing Orbit through Resilient Global Traffic Management

Philippe Goffin[1] and Gazmend Huskaj[2, 3]
[1]Military Advisor, Permanent Representation of Belgium to the United Nations, New York, USA
[2]Geneva Centre for Security Policy (GCSP), Geneva, Switzerland
[3]Department of Computer and Systems Sciences, Stockholm University, Sweden

philippe.goffin@mil.be
g.huskaj@gcsp.ch

Abstract: This paper analyses the effectiveness of the international legal framework in mitigating the combined risks of orbital collision and cyber interference, and assesses how a global Space Traffic Management (STM) regime could enhance stability and security in outer space. Using Yin's single-case study design, the global space governance regime is examined through Bowen's qualitative document analysis of treaties, policies, and twelve scholarly studies, interpreted via Dunn's public-policy analysis framework. The findings show that existing instruments establish broad norms of responsibility and liability but lack operational authority, cyber-resilient procedures, and enforceable "rules of the road." Fragmented institutions and divergent national strategies perpetuate legal ambiguity and impede coordinated manoeuvre decisions. Cyber dependencies in orbital systems further magnify risks of misinterpretation and escalation. The study argues that an integrated STM regime—combining binding collision-avoidance authority, mandatory cyber-security standards, and interoperable data architectures—would materially reduce uncertainty and strengthen collective security. Policy recommendations include establishing a supranational coordination centre, codifying minimum separation and priority rules, and linking participation to verifiable cyber compliance. The proposed framework operationalises due regard through authority, assurance, and accountability, transforming space governance from permissive norms to enforceable safety mechanisms. By fusing legal, institutional, and technological dimensions, the research provides a practical pathway toward a secure and predictable orbital environment.

Keywords: Space governance, Cyber resilience, Cyber warfare, Space Traffic Management, Artificial intelligence, International law

1. Introduction

Space underpins contemporary national security, economic activity, and societal functioning by enabling communications, navigation, remote sensing, and intelligence across civil and military domains. The rapid expansion of orbital activity has, however, transformed once-predictable operating conditions into a congested and increasingly fragile environment. By 2024, space-surveillance networks tracked more than 10,000 operational satellites and over 36,000 catalogued objects, while commercial constellations such as Starlink executed tens of thousands of collision-avoidance manoeuvres annually to maintain safe separation (ESA, 2024; SpaceX, 2023). This accelerated growth—combined with uneven data sharing and the absence of unified coordination—creates conditions in which a technical malfunction, missed notification, or lost control link may be misinterpreted as intentional interference, thereby elevating escalation risks in an already volatile geopolitical landscape (Johnson, 2009).

Although early legal instruments, including the Outer Space Treaty (United Nations, 1967) and the Liability Convention (United Nations, 1972), establish principles of responsibility, liability, and due regard, they were drafted for an era of limited actors and stable trajectories. They do not reflect today's orbital environment, which is characterised by large commercial constellations, rapid manoeuvre cycles, and software-defined satellite architectures. Modern space systems rely on digitally interconnected command links, cloud-integrated ground segments, and update-driven control software. These dependencies expose orbital operations to cyber intrusions that can degrade satellite functionality, distort telemetry, or delay manoeuvres—producing physical and strategic consequences that reverberate across domains (Hallaq, 2017; Birk & Kovács, 2019; Hsu & Falco, 2023; Lykou et al., 2023). The combined effect is a dual risk: collision hazards arising from congestion and cyber-enabled disruptions that can obstruct manoeuvre execution or lead to misinterpretation of intent.

Against this backdrop, the core problem examined in this paper is twofold. First, the current international legal framework provides broad normative guidance but lacks binding, time-sensitive mechanisms for coordinated collision avoidance and digital incident response. Second, the absence of a cyber-resilient governance structure increases the likelihood that routine orbital events—whether technical faults, cyber interference, or sensor anomalies—could cascade into diplomatic disputes or security crises. This study therefore assesses the adequacy of existing legal and institutional arrangements in mitigating these risks and evaluates how a global

Space Traffic Management (STM) regime could integrate authoritative coordination, cyber-resilience, and interoperable data architectures.

The research question guiding the analysis is: *To what extent does the existing international legal framework mitigate the dual risks of orbital collision and cyber interference, and how might a global Space Traffic Management (STM) regime enhance stability and security in outer space?*

The article proceeds as follows. Section 2 reviews the scientific literature on space governance, cyber interdependencies, and emerging AI-enabled operational concepts. Section 3 outlines the methodological design. Section 4 presents empirical findings derived from qualitative document analysis. Section 5 develops the policy discussion and identifies implications for constructing an integrated and enforceable STM framework.

2. Reviewing the Scientific Literature

Research on space governance converges on the view that expanding satellite constellations and increasing debris densities have intensified concerns about congestion and coordination, prompting calls for more structured mechanisms to manage orbital interactions (ESA, 2024; ITU, 2022; SpaceX, 2023). Existing Space Situational Awareness (SSA) and Space Surveillance and Tracking (SST) arrangements are widely assessed as fragmented and voluntary, with scholars arguing that the absence of a global authority for manoeuvre coordination or interference attribution creates persistent operational and geopolitical vulnerabilities (EU High Representative, 2023; Stoltenberg, 2016).

Parallel scholarship extends these concerns into the cyber domain. Studies of satellite infrastructures emphasise their dependence on software-defined components, networked command links, and ground-segment connectivity, which collectively expose orbital systems to manipulation and cascading failure (Henselmann & Lehto, 2019; Birk & Kovács, 2019). Analyses of cyber operations against space assets argue that such activities can generate kinetic and strategic effects, blurring the boundary between terrestrial and orbital conflict (Hsu & Falco, 2023; Coppock et al., 2024). Research on ontology-based situational awareness models advances proposals for automated reasoning and early detection within space–cyber ecosystems, linking technical precision to operational resilience.

Artificial intelligence is increasingly positioned as a driver of this interdependence. Hallaq et al. (2017) examine machine learning as an enabler of military decision-making, noting that greater autonomy in detection and response may reduce human oversight and heighten escalation risk. Segate (2024) similarly argues that automation in space–cyberspace operations requires updated norms of attribution and accountability. Legal scholars highlight that the existing governance framework remains misaligned with these digital dynamics, lacking mechanisms for enforcement or coordinated response (Birk, 2021; Birk & Kovács, 2019; Stefoudi, 2019). Dilworth and Osborne (2022) add that although the Outer Space Treaty's responsibility and liability provisions may apply to cyber operations, attribution challenges and weak enforcement limit their practical utility during time-critical incidents. Gompert and Libicki (2023) complement these arguments by framing "detect and engage" as an emerging operational paradigm in which AI-enabled sensing, distributed space systems, and cyber operations accelerate decision cycles across domains.

Parallel Chinese research reinforces these themes from a technical perspective. Liu et al. (2023) propose ontology-based situational awareness models for space–cyber threats, complementing legal and institutional analyses of deterrence. Across this scholarship, three insights recur: (1) the legal inadequacy of existing space governance instruments to address cyber-enabled risks; (2) the deepening technical interdependence between orbital and digital infrastructures; and (3) an emerging consensus that a cyber-resilient Space Traffic Management (STM) regime is needed to integrate ontological awareness, AI-assisted defence, and enforceable coordination.

3. Methods and Materials

This study adopts a mono-method qualitative research design consistent with Yin's (2018) single-case study approach. The "case" under examination is defined as the global space governance regime—a bounded system comprising international treaties, national policies, and institutional frameworks governing activities in outer space. This case was selected due to its central relevance to the research problem: the inadequacy of existing legal and institutional mechanisms to manage risks arising from the intersection of space traffic, cyberspace operations, and international security. The single-case design enables an in-depth contextual analysis of this

governance structure within its real-world setting, where the boundaries between legal, technological, and strategic dimensions are interdependent.

Data collection followed the logic of triangulated documentary evidence, drawing from two primary sources. First, a structured literature review was conducted using the Scopus database to identify peer-reviewed publications addressing cyberspace operations, satellite vulnerabilities, and cybersecurity governance between 2015 and 2025. Boolean operators:

> *{satellite} AND ("cyber operation*" OR "offensive cyber*" OR "cyber warfare" OR "cyber conflict*" OR "cyber defence" OR "cyber defense") AND ("satellite*" OR "space system*" OR "orbital" OR "space asset*" OR "spacecraft") OR ((Cyber* AND "vulnerab*") AND ("satellite communication*" OR "SATCOM" OR "space network*" OR "ground station*")) OR (("space traffic management" OR "STM" OR "space governance" OR "outer space law" OR "international law") AND ("cyber secur*" OR "cyber norm*" OR "cyber governance" OR "digital resilience" OR "information assurance"))*

were applied to capture the conceptual breadth of the field. The query initially yielded 119 results; after refinement and screening based on relevance, citation frequency, and data accessibility, a final corpus of twelve academic articles was retained for detailed review. Second, primary legal and institutional documents—including United Nations treaties, European Union policies, national space directives, and official reports—were examined to provide the empirical foundation for the analysis.

Data analysis employed Qualitative Document Analysis (QDA) as formulated by Bowen (2009). This involved systematic skimming, detailed reading, and iterative interpretation of the selected materials to identify patterns, principles, and governance gaps. Analytical emphasis was placed on tracing how international law and institutional mechanisms address, or fail to address, cyber–space interdependencies. Dunn's (2018) public policy analysis framework was applied as the interpretive lens to translate these empirical findings into policy insights and recommendations. Through this integrated approach, the study aligns Yin's contextual depth with Bowen's methodological discipline and Dunn's analytical reasoning, generating a comprehensive understanding of the structural, legal, and normative dimensions of global space governance.

4. Results

This section presents the findings of a qualitative document analysis (QDA) of the current global space governance regime as it intersects with cyberspace operations and space traffic management. The corpus comprised foundational United Nations treaties and principles, recent UN policy texts, regional and national policy artefacts, and twelve peer-reviewed sources addressing cyberspace operations, AI-enabled defence, and space governance. The analysis identifies recurring categories that structure the empirical record: the evolution of governance instruments; emerging cyber dependencies in orbital systems; institutional fragmentation and legal ambiguity; national and regional policy divergence; and the technological integration and automation dilemma. Findings are descriptive and evidence-based, reporting what the documents contain and how patterns recur across sources.

4.1 The Evolution of Space Governance Instruments

Across the legal corpus, the baseline architecture for space activities is stable in its core obligations yet sparse in operational detail. The Outer Space Treaty establishes responsibility, liability, and due regard duties, with States bearing international responsibility for national activities, including those of non-governmental entities (United Nations, 1967). The Liability Convention further specifies compensation for damage caused by space objects, while the Registration Convention obliges delivery of orbital data to a central registry (United Nations, 1972a; United Nations, 1972b). In the analysed texts, these instruments provide a permissive legal envelope rather than granular navigation rules. The treaty language contains no codified priorities for "right-of-way," minimum separation standards, or binding collision avoidance procedures applicable to routine on-orbit operations outside manoeuvre and launch contexts (United Nations, 1967). The Agreement Governing Activities on the Moon contains more specific coordination references but confines them to lunar environments (United Nations, 1979).

Soft-law sources bridge some gaps but remain non-binding. The 1964 Declaration reaffirms State responsibility and enables consultations where potentially harmful interference is foreseen, yet it leaves initiative, timelines, and decision authority undefined (UN General Assembly, 1964). Subsequent principles on direct broadcasting, remote sensing, nuclear power sources, and international cooperation add sectoral process guidance and ITU-mediated frequency coordination but do not institute live orbital traffic deconfliction mechanisms (UN General

Assembly, 1983; 1987; 1993; 1997). Recent UN system texts—such as the Guidelines for the Long-Term Sustainability (LTS) of Outer Space Activities—articulate best practices for mission planning, data sharing, and risk reduction, without establishing an enforcement locus for trajectory conflict resolution (UN OOSA, 2021).

The "Pact for the Future" identifies discussion tracks for space traffic, debris, and resources, thereby signalling agenda-setting rather than an operational governance shift (UN General Assembly, 2024). Taken together, the instruments exhibit continuity in high-level norms and persistent absence of prescriptive "rules of the road," a pattern that underpins subsequent empirical categories concerning cyber-physical integration and traffic management interfaces (United Nations, 1967; 1972a; 1972b; 1979; UN OOSA, 2021; UN General Assembly, 2024).

4.2 Emerging Cyber Dependencies in Orbital Systems

Across the technical and policy literature, orbital infrastructures are increasingly co-constituted with cyber capabilities that enable, expose, and configure operational behaviours. Peer-reviewed studies describe satellite networks as networked information systems reliant on protocol stacks, ground segment connectivity, and cloud-mediated services, with vulnerabilities spanning denial-of-service, malicious code insertion, and data exfiltration across RF and IP layers (Aviv & Ferri, 2023; She et al., 2017; Xu, 2024). Analysis of moving-target defence (MTD) in satellite internet prototypes demonstrates that pseudo-random address and port hopping can increase attacker cost without significant throughput penalties, indicating a direction of travel toward dynamic cyber manoeuvres at the network edge of orbital architectures (Xu, 2024). Legal scholarship reveals that cyberspace operations against space systems can manifest as interference with command links, data links, or payload integrity, raising attribution and remedy questions under existing space and general international law (Schmitt & Vihul, 2017).

Empirical reporting on the Russia–Ukraine armed conflict provides concrete instances of cyber–space interdependence. Wiper operations, DDoS campaigns, and the Viasat KA-SAT incident illustrate how cyber effects can degrade space-enabled services, while commercial LEO broadband constellations supplied resilient communications and positioning-navigation-timing (PNT) workarounds through software updates and constellation-level redundancy (Aviv & Ferri, 2023). The same sources document accelerated migration of governmental data from damaged on-premises facilities to public cloud platforms and partner-hosted private clouds outside the national territory, thereby re-situating space-derived data flows within distributed digital infrastructure subject to varied legal jurisdictions (Aviv & Ferri, 2023).

At the interface of cyber and electromagnetic operations, analyses of spectrum-centred tactics describe how jamming, spoofing, and cyber payloads co-evolve, with Intelligence, Surveillance, and Reconnaissance (ISR) systems and command-and-control relying on assured spectrum access and cyber-hardened communication links (Young, 2017). Across documents, the empirical pattern is a deepening reliance on software-defined, cloud-connected, and update-driven control surfaces for orbital systems, which couples cyber resilience properties to space traffic safety outcomes via command availability, data integrity, and timing assurance (Aviv & Ferri, 2023; She et al., 2017; Xu, 2024; Young, 2017). This dependency anchors subsequent findings on the adequacy of institutional arrangements to accommodate cyber-physical risk within traffic management.

4.3 Institutional Fragmentation and Legal Ambiguity

The document set consistently records a dispersion of roles across technical standard-setters, national regulators, and international bodies, with limited consolidation around an entity empowered to issue binding orbital deconfliction instructions. COPUOS processes, including the LTS Guidelines, emphasise voluntary practices and information exchange, while ITU coordination focuses on frequencies and orbital slots, not operational manoeuvre authority for collision risk events (UN OOSA, 2021). Legal analyses highlight that, while the liability and responsibility framework supplies post-hoc remedies and State-to-State accountability, it does not specify pre-emption thresholds, procedural priorities, or operational control for time-sensitive conjunctions involving mixed public–private constellations (Schmitt & Vihul, 2017). Recent UN policy texts call for exploring frameworks on traffic and debris, yet they stop short of allocating decision rights or establishing supranational service provision for space traffic management (UN General Assembly, 2024).

Parallel institutional initiatives aim to fill cyber-specific gaps but remain emergent. Drafting efforts for cybersecurity standards tailored to outer space missions and critical infrastructure—surveyed in the technical governance literature—map candidate control baselines and assurance processes, often drawing from terrestrial frameworks and CCSDS security work, without an associated compliance jurisdiction in orbit (Segate, 2024). Concept papers on "assured cyber deterrence" in space document doctrinal debates around

"hacking back," malware pre-positioning, and "booby traps," illustrating how proposed cyber response logics intersect with the neutral-use and due-regard obligations of space law in the absence of authoritative interpretive guidance (Ting & Lin, 2020). The accumulated materials therefore show an institutional landscape where voluntary guidelines, sectoral standards, and national practices co-exist, but where neither treaty text nor standing body assigns operational authority for collision avoidance when cyber availability or integrity constraints affect manoeuvre timelines. This fragmentation conditions the national policy divergence recorded in the next category.

4.4 National and Regional Policy Divergence

Policy texts reveal differing national and regional approaches to space traffic management and cyber–space risk governance. The United States' 2018 Space Traffic Management Policy identifies flight-safety challenges and positions national leadership, industry solutions, and best-practice diffusion as primary levers, with limited appetite for delegating traffic control to multilateral entities (White House, 2018). European materials in the corpus describe multiple SSA/SST initiatives and data-sharing arrangements supported by the EU and ESA, yet note the absence of a comprehensive STM policy allocating coordination or control for avoidance manoeuvres across Member States (EU High Representative, 2023; ESA, 2024). The reviewed texts characterise China's posture through technical and standards-oriented publications and national capability development, with growing attention to active cyber defence at network and platform layers in satellite internet contexts (The State Council of the People's Republic of China, 2021; Xu, 2024; Segate, 2024). In the UN arena, the "Pact for the Future" places STM, debris, and resources on the intergovernmental agenda and invites private sector participation, but does not harmonise State preferences regarding locus of control (UN General Assembly, 2024).

Empirical accounts from the Russia–Ukraine conflict show how national strategies shape cyber-resilient operations using public–private assemblages. Ukraine's migration of sensitive registries to public clouds and to allied-hosted private infrastructure, combined with the adoption of commercial LEO broadband for continuity under kinetic and cyber stress, depict a resilience pathway grounded in diversified jurisdictional and commercial arrangements rather than a unified international traffic authority (Aviv & Ferri, 2023). Regional telecom measures, including internal roaming and inbound roaming blocks, further illustrate how national instruments condition service continuity and exposure in crisis, with secondary effects on spectrum and network management for satellite-backed services (Aviv & Ferri, 2023). The materials thus indicate variation in governance choices—lead-nation versus multilateral coordination, public cloud reliance versus on-premises control—that collectively produce a patchwork of practices relevant to traffic safety when cyber availability determines manoeuvre execution windows.

4.5 Technological Integration and the AI–Automation Dilemma

Across defence, security, and engineering sources, AI-enabled functions are increasingly embedded in space and cyberspace operations, raising documentable questions about control allocation, assurance, and failure modes. Reviews of AI in military cyber contexts describe learning-based detection, decision support, and autonomy for ISR, command-and-control, and defensive cyber manoeuvre, situating these capabilities within dual-use trajectories (Hallaq, 2017). Maritime cybersecurity research on dual-use AI offers parallel empirical observations on automated anomaly detection, adversarial adaptation, and model governance, which are transferable to spaceborne and ground segment networks that share protocol and operations patterns (Aldhaheri et al., 2020). Operational concepts emphasising "detect and engage" in contested electromagnetic and cyber environments report increased dependence on rapid machine-assisted sense-making and decision cycles, with electromagnetic control and cyber persistence viewed as central to force employment (Deptula & Bacon, 2019; Young, 2017).

Within the space domain specifically, security scholarship documents proposals for pre-positioned code, active countermeasures, and automated responses to cyber intrusion on spacecraft and ground segment systems, noting potential intersections with space law obligations in the absence of agreed interpretations (Ting & Lin, 2020; Schmitt & Vihul, 2017). Observational materials from the Russia–Ukraine conflict capture software-based mitigation at constellation scale, including over-the-air code changes to counter jamming and sustain PNT-related services, underscoring the operational reality of software-defined remediation in orbit (Aviv & Ferri, 2023).

The technical governance literature registers contemporaneous efforts to draft cybersecurity standards for space missions—threat models, control families, and assurance artefacts—without a clear mapping to traffic

management decisions when automated cyber responses affect attitude control, propulsion commands, or conjunction screening timelines (Segate, 2024). The pattern across documents is an integration of AI and automation into cyber-space operations that improves detection and continuity while complicating accountability chains and temporal coupling with STM processes. This linkage provides the empirical basis for evaluating adequacy of current governance in subsequent discussion. Table 1 summarises the core governance gaps identified across the legal, institutional, cyber, and technological dimensions.

Table 1: Summary of Governance Gaps

Dimension	Key Gap Identified	Evidence Sources	Implication for STM
Legal	No "rules of the road," unclear priority rights	OST 1967, Liability Convention 1972	No binding authority for time-critical manoeuvres
Institutional	Fragmented roles, voluntary data sharing	COPUOS LTS, ITU regulations	No central coordination for deconfliction
Cyber	Vulnerable command links, no cyber baselines	Birk & Kovács 2019; Hsu & Falco 2023	Cyber interference can delay or distort manoeuvres
Technological	AI automation without governance	Hallaq 2017; Segate 2024	Machine-speed decisions create accountability gaps

5. Discussion

The answer to the research question, To what extent does the existing international legal framework mitigate the dual risks of orbital collision and cyber interference, and how might a global Space Traffic Management (STM) regime enhance stability and security in outer space?, is that current instruments partially shape responsible conduct yet fail to govern time-critical deconfliction or cyber-physical interdependence. The present framework constrains overtly harmful behaviour and clarifies post-hoc responsibility, but it does not allocate decision rights for live conjunction management or define cyber-resilient operating procedures. A global STM regime that fuses authoritative traffic coordination with mandatory incident reporting, interoperable data exchange, and baseline cyber hygiene would materially reduce collision likelihood, dampen misinterpretation, and enhance strategic stability. In practice, progress depends on aligning legal, institutional, and technological arrangements to support predictable manoeuvre decisions under conditions of uncertainty and time pressure.

The results reveal structural weaknesses in law, authority, and technology. Current regulations define responsibility and liability after incidents but do not specify who has priority, what distance is safe, or how manoeuvre decisions are made. Institutions promote cooperation and information exchange but lack the authority to issue binding instructions when collision risks arise. Satellites depend on software-based control, cloud connectivity, and constant updates, exposing them to cyber disruptions that delay or distort manoeuvres. The problem is systemic: rules exist without enforcement, coordination mechanisms without command authority, and safety processes that rely on unstable communication links. Without integration of these elements, routine events in orbit can be misread as deliberate acts, generating disputes over intent and responsibility.

The data also indicate growing collision risks, increased automation, and greater uncertainty about intent. Expanding constellations compress decision time and require automated systems to filter and prioritise manoeuvres. As autonomy advances, actions occur at machine speed, reducing transparency and complicating verification. Cyber operations targeting ground systems or command links can disrupt control and appear as deliberate obstruction. National approaches will continue to diverge, with some favouring market solutions and others multilateral coordination. Without a shared baseline for space-traffic management, the orbital environment will face higher near-miss rates, inconsistent manoeuvre policies, and restricted data exchange that weakens collective awareness. The main risk is not intentional attack but misjudged behaviour, where technical failure or interference is read as aggression under stress.

Policy measures should aim for steady progress that links traffic authority to cyber resilience. A coordination function should be established with the power to issue time-bounded avoidance directives under transparent procedures and review mechanisms. Clear and technology-neutral operating rules are needed, defining separation thresholds by orbital regime, priority based on mission and manoeuvre capability, and agreed timelines for response. Participation must require baseline cyber controls, including authentication of command links, integrity protection for tracking data, and timely disclosure of incidents. A federated data structure should merge public and commercial sources under verifiable quality standards and auditable access.

Incentives should align through insurance discounts for compliance, launch licensing tied to adherence, and procurement preferences for transparent data-sharing. These reforms must balance national sovereignty, commercial confidentiality, and the cost of upgrading existing systems.

Monitoring must translate these reforms into measurable indicators. Safety metrics should include collision probabilities, near-miss counts, and average delay from alert to manoeuvre. Cyber-resilience metrics should track link availability, verified data integrity, and the number of cyber incidents that impede operations. Governance metrics should measure data convergence across providers, participation rates, compliance with directives, and the timeliness of reporting. Capability metrics should record operators with verified contingency modes, independent verification coverage, and the frequency of joint exercises. Together, these indicators enable adaptive learning, targeted support, and proportionate responses when safety thresholds are crossed.

Evaluation criteria should be defined in advance to maintain focus and prevent goal drift. Effectiveness should test whether collision risk and ambiguity about intent decline over time. Efficiency should assess whether coordination costs are justified by fewer claims, outages, and debris growth. Equity should examine the effects on new entrants, ensuring fair access and manageable obligations. Legality and legitimacy must consider consistency with treaties, transparency in decision-making, and due-process safeguards. Resilience should test whether the regime withstands cyber or sensor failures without cascading disruption. Evidence should include baseline-to-outcome comparisons, post-incident analysis, and red-team testing of automated decision tools. Where performance fails to meet standards, corrective clauses should trigger procedural or technical revision. Feasibility depends on institutional design, decision rights, and sequencing.

A treaty-anchored framework should establish a coordination centre with a narrow mandate to issue avoidance directives under defined conditions. Broader participation can follow through mutual recognition of national licences embedding STM and cyber requirements. Interoperability should rely on open technical standards for data formats, encryption, and interface testing. Funding may combine assessed contributions for shared functions, user fees for premium services, and penalties for non-compliance that finance remediation. Disputes should be resolved through two tracks: fast-track technical adjudication for manoeuvre conflicts and a slower legal process for liability claims. This sequencing lowers entry barriers while maintaining a clear path toward enforceable coordination.

Risk management must address automation failures, malicious behaviour, and compromised data. Autonomy must include human-in-the-loop thresholds, verified fallback modes, and independent validation of critical decisions. Agreed “hold” protocols should define when to suspend manoeuvres during catalogue errors or cyber incidents. Data exchanges must follow zero-trust principles with continuous verification of ground and onboard systems and cryptographic provenance for trajectory data. Electromagnetic protection should complement cyber controls, as jamming and spoofing often coincide with data manipulation. Confidence-building should include direct communication channels between operators, post-incident reviews, and anonymised reporting to improve learning and accountability.

An effective traffic-management system must also accommodate commercial activity and innovation. Data-trust models can support analysis of sensitive orbital data without disclosing proprietary details. Tiered access can separate safety-critical information from competitive services. Regulatory sandboxes can test automation in controlled environments, expanding only after proven reliability. Support for emerging spacefaring nations should include tools, training, and subsidised certification to ensure equitable participation and reduce systemic risk from uneven capacity. Over time, better data, clearer rules, and trusted automation reinforce each other—improving safety and preserving accountability.

The implications for governance and international security are clear. A system that integrates authority, assurance, and accountability can turn a permissive legal framework into an operational safety regime. Coupling traffic management with secure infrastructures and measurable performance strengthens trust, narrows uncertainty, and limits escalation. This approach does not replace existing treaties or national interests; it translates due-regard obligations into shared procedures, verified data, and coordinated action. The discussion thus prepares the ground for final policy conclusions, a phased implementation roadmap, and further research on human–machine coordination, liability in automated decisions, and the governance of adaptive systems that will shape the future of orbital safety.

Acknowledgements

Artificial intelligence tools were used only to assist with research organisation and language editing. All analysis, interpretation, and conclusions are the authors' own. Consistent with Nature's (2023) guidelines, AI tools are not credited as authors.

Ethical declaration: The study involved no human participants or sensitive data, so ethical approval was not required.

Disclaimer: The views expressed in this paper are solely those of the authors and do not necessarily reflect the official policies or positions of the Belgian Armed Forces, the Government of Belgium, the Geneva Centre for Security Policy, or Stockholm University. The analysis and conclusions are presented for academic and policy discussion purposes only.

References

Aviv, I., & Ferri, U. (2023). *Russian–Ukraine armed conflict: Lessons learned on the digital ecosystem.* International Journal of Critical Infrastructure Protection, 43, 100637. https://doi.org/10.1016/j.ijcip.2023.100637

Birk, A. M. (2021). Legal and policy aspects of satellite and space traffic management. *Space Policy, 57,* 101420.

Birk, A. M., & Kovács, L. (2019). Cyber operations in satellite infrastructures: Emerging vulnerabilities and policy responses. *Journal of Cybersecurity, 5*(3), tyz013.

Bowen, G. A. (2009). *Document analysis as a qualitative research method. Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027

Coppock, A., Jensen, R., & Lavigne, D. (2024). Artificial intelligence in maritime cybersecurity: Dual-use applications for defence and offence in the age of digital seas. *Defence Studies, 24*(2), 198–215.

Dilworth, S. W., & Osborne, D. D. (2022). *Cyber threats against and in the space domain: Legal remedies*. In T. Jančárková, G. Visky, & I. Winther (Eds.), 2022 14th International Conference on Cyber Conflict: Keep Moving (pp. 235–247). NATO Cooperative Cyber Defence Centre of Excellence.

European Space Agency. (2024). *Space environment statistics · Space debris user portal.* https://sdup.esoc.esa.int/discosweb/statistics/

European Union High Representative for Foreign Affairs and Security Policy. (2023). *European Union space strategy for security and defence.* https://ec.europa.eu/transparency/documents-register/detail?ref=JOIN(2023)9&lang=en

Gompert, D. C., & Libicki, M. (2023). Detect and engage: A new American way of war. *Survival*, 65(5), 65–74. https://doi.org/10.1080/00396338.2023.2261246.

Hallaq, B., Somer, T., Osula, A.-M., Ngo, K., & Mitchener-Nissen, T. (2017). Artificial intelligence within the military domain and cyber warfare. *Journal of Information Warfare, 16*(4), 55–68. https://wrap.warwick.ac.uk/id/eprint/104774

Henselmann, G., & Lehto, M. (2019). Where cyber meets the electromagnetic spectrum. In T. Cruz & P. Simões (Eds.), *Proceedings of the 18th European Conference on Cyber Warfare and Security (ECCWS 2019)* (pp. 209–218). Academic Conferences International.

Hsu, J., & Falco, G. (2023). *Space booby traps: Hacking back and assured cyber deterrence in space.* In *2023 IEEE International Conference on Assured Autonomy (ICAA)* (pp. 115–118). IEEE. https://doi.org/10.1109/ICAA58325.2023.00024.

International Telecommunication Union. (2022). *E-submission of satellite network filings.* https://www.itu.int/ITU-R/space/asreceived/Publication/DisplayPublication/23706

Johnson, N. (2009). *The collision of Iridium 33 and Cosmos 2251: The shape of things to come.* NASA Technical Reports Server. https://ntrs.nasa.gov/api/citations/20100002023/downloads/20100002023.pdf

Kovács, L., & Tóth, K. (2019). Cyber threats to space systems and implications for national security. *Journal of Strategic Security, 12*(4), 75–90.

Liu, B., Yi, J., Yao, L., Wang, Y., Ding, Z., & Zhu, X. (2023). 面向太空网络战威胁的态势感知本体建模 [Ontology-based situational awareness for space cyber threats]. *Journal of Cybersecurity Science and Technology, 8*(1), 33–46.

Lykou, S., Gritzalis, D., & Theocharidou, M. (2023). Cybersecurity of space systems and satellite communication infrastructures. *Acta Astronautica, 206,* 141–156.

Nature Editorial. (2023). *Tools such as ChatGPT threaten transparent science; here are our ground rules for their use*. *Nature*, 613, 612. https://doi.org/10.1038/d41586-023-00191-1

Stefoudi, D. (2019). *The relevance and applicability of cybersecurity laws with regard to data storage on board satellites and on the ground*. Air & Space Law, 44(4–5), 425–444.

SpaceX. (2023). *SpaceX Gen 1 semi-annual report.* https://licensing.fcc.gov/myibfs/download.do?attachment_key=23204338

Stoltenberg, J. (2016). *Press conference following the meeting of the North Atlantic Council at the level of Heads of State and/or Government.* North Atlantic Treaty Organization. https://www.nato.int/cps/en/natohq/opinions_171554.htm

The State Council of the People's Republic of China. (2021). *China's space program: A 2021 perspective.* https://english.www.gov.cn/archive/whitepaper/202201/28/content_WS61f35b3dc6d09c94e48a467a.html

UN General Assembly. (2024). *Pact for the Future, Global Digital Compact and Declaration on Future Generations.* https://www.un.org/sites/un2.un.org/files/sotf-pact_for_the_future_adopted.pdf

United Nations General Assembly. (1964). *Declaration of legal principles governing the activities of states in the exploration and use of outer space (18th session).* United Nations Digital Library. https://digitallibrary.un.org/record/203965

United Nations General Assembly. (1983). *Principles governing the use by states of artificial Earth satellites for international direct television broadcasting (37th session).* United Nations Digital Library. https://digitallibrary.un.org/record/41084

United Nations General Assembly. (1987). *Principles relating to remote sensing of the Earth from outer space (41st session).* United Nations Digital Library. https://digitallibrary.un.org/record/126423

United Nations General Assembly. (1993a). *Declaration on international cooperation in the exploration and use of outer space for the benefit and in the interest of all states, taking into particular account the needs of developing countries (51st session).* United Nations Digital Library. https://digitallibrary.un.org/record/231739

United Nations General Assembly. (1993b). *Principles relevant to the use of nuclear power sources in outer space (51st session).* United Nations Digital Library. https://digitallibrary.un.org/record/159141

United Nations Office for Outer Space Affairs. (2021). *Guidelines for the long-term sustainability of outer space activities of the Committee on the Peaceful Uses of Outer Space.* https://www.unoosa.org/documents/pdf/PromotingSpaceSustainability/Publication_Final_English_June2021.pdf

United Nations. (1967). *Treaty on principles governing the activities of states in the exploration and use of outer space, including the Moon and other celestial bodies.* United Nations Treaty Series. https://treaties.un.org/doc/Publication/UNTS/Volume_610/volume-610-I-8843-English.pdf

United Nations. (1972a). *Convention on registration of objects launched into outer space.* United Nations Treaty Series, 1023(1), 15020. https://treaties.un.org/doc/Publication/UNTS/Volume_1023/volume-1023-I-15020-English.pdf

United Nations. (1972b). *Convention on the international liability for damage caused by space objects.* United Nations Treaty Series, 961(1), 13810. https://treaties.un.org/doc/Publication/UNTS/Volume_961/volume-961-I-13810-English.pdf

United Nations. (1979). *Agreement governing the activities of states on the Moon and other celestial bodies.* United Nations Treaty Series, 1363, 23002. https://treaties.un.org/doc/Publication/UNTS/Volume_1363/volume-1363-I-23002-English.pdf

Vecellio Segate, R. (2024). Drafting a cybersecurity standard for outer-space missions on critical infrastructure: China and the West. *Computers & Security, 138,* 103749. https://doi.org/10.1016/j.cose.2024.103749

White House. (2018). *Space Policy Directive-3: National Space Traffic Management Policy.* https://trumpwhitehouse.archives.gov/presidential-actions/space-policy-directive-3-national-space-traffic-management-policy/

Xu, R. (2024). 基于移动目标防御的卫星互联网防御系统设计 [Design of satellite Internet defense system based on mobile target defense]. *Space-Integrated-Ground Information Networks, 5*(4), 107–112. https://doi.org/10.11959/j.issn.2096-8930.2024043.

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). Thousand Oaks, CA: SAGE Publications.

Ontological Security and Threat Mitigation in Healthcare: A Descriptive Exploration

Brandon Griffin[1], Michaela Barnett[2], Tom VanNorman[3], Nina Janine Medina[3], Lucas Potter[4] and Xavier Palmer[4]

[1]Cyberlinc, USA
[2]Blacks In Cybersecurity Headquarters, Inc., USA
[3]Biohacking Village, USA
[4]BIOSView Labs, USA

bgrif023@gmail.com
michaela@bichq.org

Abstract: The healthcare sector's growing reliance on interconnected cyber-physical systems, from electronic health records (EHRs) to networked medical devices, has expanded both operational capabilities and systemic vulnerabilities. The costs associated with cyber-attacks have become a financial burden for the global healthcare sector, with estimates whose substance is capable of bankrupting institutions. These costs underscore the urgent need to analyze expanding attack surfaces at the intersection of cybersecurity and biosecurity [cyberbiosecurity (CBS)/biocybersecurity (BCS)], requiring systematic identification of targetable assets and corresponding defense protocols. This paper analyzes healthcare's evolving threat landscape through the lens of ontological security. The commentary focuses specifically on threats to living systems and biological data integrity (e.g., genetic, biometric), suggesting how a focus on these unique biological targets necessitates a shift from purely IT-centric defense to a human-centric risk posture. Addressing this, we draw on the synthetic definition of ontological security which addresses stability from continuity of experience in one's life and apply it to the healthcare space. From here, we build on the notion of perceptions of secure healthcare spaces aiding care received, we attempt to spotlight potential consequences relating to vulnerabilities in healthcare that could link to lapses in patient care, including identity fragmentation during biometric theft, temporal collapse from ransomware-induced care delays, and spatial destabilization via telemedicine breaches. The paper provides a commentary that draws on select literature, drawing on a combination of techniques and observations that can restore lapses in ontological security. We address ethical imperatives for protecting biometric and genetic data, as central targets in this expanded threat landscape, while outlining policy measures to future-proof healthcare systems against AI-driven threats. By centering existential safety alongside technical safeguards, this work redefines healthcare cybersecurity as a covenant of human security in an increasingly digitized care ecosystem. The goal of this exploration is to provide global health institutions with considerations to aid care strategies that simultaneously protect patient safety, provider integrity, and institutional resilience.

Keywords: Ontological security, Cybersecurity, Biocybersecurity, Cyberbiosecurity, Attack surfaces, Healthcare

1. Introduction

Attacks on healthcare are part of the "new normal" within cybersecurity, and failure to account for the consequences of actors exploiting these attack surfaces, especially with advanced technologies, can yield severe consequences (Razaque et al, 2019; Dimitrov, 2020; Fouad, 2024). Modern healthcare operates within an increasingly interconnected cyber-physical ecosystem where critical infrastructure, including electronic health records (EHRs) systems, networked medical devices and telemedicine platforms converge to enhance patient care. However, this digitization expands the attack surface for malicious actors, introducing unprecedented risk to data integrity, service continuity and ontological security. "Attack surfaces" (Theisen et al, 2018) as a term encompasses numerous definitions surrounding the context and purposes of this paper. We refer to numerous cyber-physical environments and surfaces that can be exploited to be the target of an attack. Ontological security is the foundational sense of safety and trust to be used in this context to refer to a Cybersecurity disposition (Griffin et al., 2023). Cyber-physical attacks have the chance to upend a sense of safety and trust, particularly within the delicate subject matter of patient care. Healthcare institutions represent target-rich environments for cyber-physical threats (Al;-Qarni, 2023; Li et al, 2025). Technical vulnerabilities in medical devices such as ventilators, pacemakers, patient monitors or infusion pumps can combine with human factors and interaction from activity such as social engineering, phishing or baiting. Additionally, supply chain weaknesses create exploitable entry points; these threats extend beyond operational disruptions to directly compromise patient ontological security. Breakdowns in institutional reliability during ransomware-induced care delays are the latest, with a notable example being the WannaCry attack (Collier, 2017). The COVID-19 pandemic significantly exacerbated these vulnerabilities by accelerating telemedicine adoption in an uncertain time (Alanazi, 2023). Additionally, inclusion of cloud computing and

adjacent technology has presented numerous attack surfaces to protect and evaluate (Kubendiran et al, 2015; Zhang et al, 2015; Van Devender et al, 2023).

This paper analyzes healthcare's attack surfaces through the lens of ontological security, examining mainly identifiable interconnected target groups; technical systems, operational technology (OT), information technology (IT), human targets, and systemic structures, among others. Initially, technical systems where compromise enables life-threatening manipulation of devices or theft of immutable biometric data. Further, human targets including clinicians, patients and supporting staff where coerced actions or eroded agency undermine relational trust. Additionally, systemic structures such as supply chains and access control where failures disrupt the environmental stability essential to ontological safety. Without enforceable standards and cross-sector collaboration, these vulnerabilities risk institutionalizing ontological insecurity. This manifests through care avoidance, clinician burnout, and general distrust (Vukotich, 2023). Effective mitigation requires socio-technical strategies addressing both cyber-physical risks and their existential repercussions. In addition to creating a framework, this work aims to open up additional considerations of security focused on BCS/CBS in the healthcare context and invite further research from adjacent communities.

2. Themes Explored

This paper offers a synthesis of healthcare infrastructure with themes or cases for analysis. Vulnerabilities across core concepts draw on a curated selection of relevant literature published between 2006 and 2024. These concepts are Cyber-physical Systems (CPS), Attack Surfaces and Vulnerabilities, Biocybersecurity (BCS) / Cyberbiosecurity (CBS), Ontological Security, Identity, Risk and Workforce Preparedness. These are thematically structured for this discussion. Pertinent literature regarding supportive frameworks was included as applicable. The selection was informed by multiple, non-systematic search sessions in relevant databases using a broad range of key terms, including, but not limited to: "cyber-physical systems," "healthcare," "cybersecurity," "Internet of Medical Things," "attack surface," "cyberbiosecurity," "biocybersecurity," "health," "biotechnology," "workforce," "defense," "ontological security," "cyber," "biotechnology," "nation-state," and "bio-economy." This commentary does not follow a systematic review methodology; rather, it acts as a commentary that highlights some important developments and connections across these fields to frame the ongoing discussion on security. BCS and CBS serve as a foundation for this commentary, having been covered extensively in prior work (Potter et al, 2021; Potter and Palmer, 2023; Greenbaum, 2023). These terms involve phenomena that have interplay between biosecurity, cyber-physical security, and cybersecurity contexts within the bioeconomy. The following sections are distillation of insights based off of such papers. These do not represent final or total thoughts on the intersections discussed but merely an opener. What is essential to note is that healthcare is not merely a technological process but one that is at the intersection of social processes, organizational dynamics, infrastructure and diverse perceptions.

3. Healthcare Attack Surfaces

Healthcare systems face a complex and overlapping set of attack surfaces that make them prime targets for cyber threats. Figure 1 gives a simple map of processes that can be made a target of. These vulnerabilities span technical, human, systemic dimensions and/or a hybrid, with each introducing unique risks to operational continuity, data integrity and ontological security.

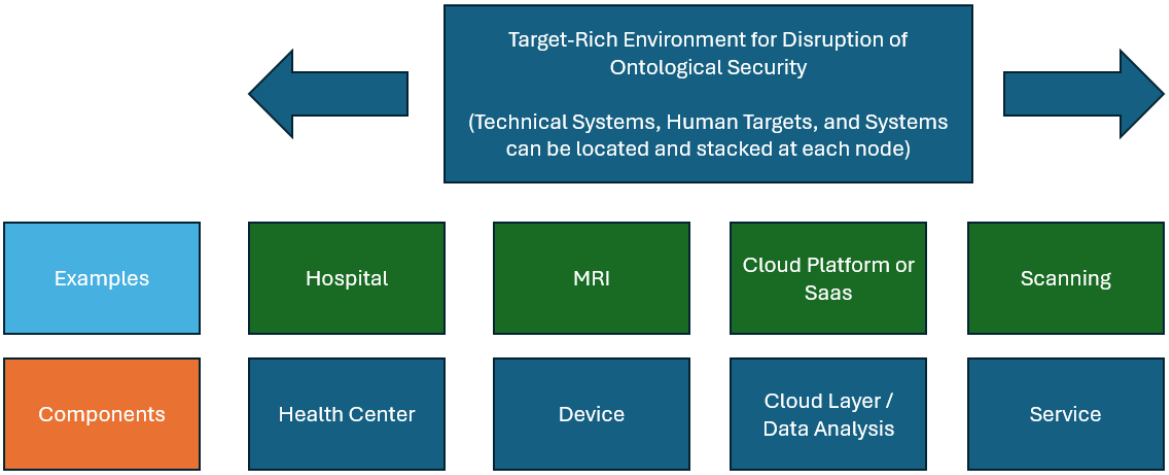


Figure 1: Map of targets within a Healthcare attack surface. Attacks can be stacked along the axes of the number of vulnerabilities and components for combinatory effects

Technical attack surfaces include devices, networks, and software, providing numerous entry points for malicious actors to exploit or disable critical systems. Examples range from networked medical devices like patient monitors, infusion pumps, or imaging modalities [Magnetic Resonance (MR), Medical Extended Reality (XR), and Computed Tomography (CT)] to electronic health record (EHR) platforms and hospital network infrastructure.

It helps planners to conceptualize the components that can be accessed by malicious elements (Spanakis et al 2020; Verma, 2022). These systems can be compromised due to device and institutional misconfigurations, weak cyber hygiene, or outdated software, and while they may not always be primary targets, their role in storing or accessing sensitive data makes them attractive secondary assets for attackers (Wasserman and Wasserman, 2022; Rajora et al, 2022; Aloufi et al, 2025). Medical devices can act as repositories for sensitive data, their role in storing patient safety information makes these a significant technical target (Mirsky, et al., 2019). Along with this, challenges and unique circumstances in asset identification and vulnerability assessment further complicate the attack surface and landscape (Skorobogatjko et al, 2014; Sultanovs et al, 2016; Van Devender & Mcdonald, 2023).

Human elements in threat modelling are often identified as involving next-of-kin, patients, staff, personnel and contractors. All of which listed are vulnerable to social engineering methods that may coerce action or initiate a malicious action. Healthcare professionals are frequently targeted due to their access to sensitive patient data and critical systems (Alavi et al., 2024; Gilbert et al, 2024). Employees such as administrative or maintenance staff, may also be exploited through financial incentives or limited cybersecurity training, providing attackers with physical or digital access to restricted areas. Their access to server rooms, storage closets, restricted areas, or personnel areas provide an advantage (Muthuppalaniappen & Stevenson, 2020). Patients and visitors represent additional targets as their personal devices and trust in healthcare environments can be manipulated to gain broader network access in various emotional, mental, or physical states or varying pain and distress levels.

Attack surfaces regarding interconnected systems, defined as “systematic” in this discussion, encompass supply chain weaknesses and auxiliary infrastructure vulnerabilities. Third-party vendors, outsourced IT services, and medical suppliers often operate with inconsistent security standards. This creates hidden entry points for attackers. Technical compromises can lead to life-threatening situations, such as manipulated medical devices or ransomware-induced treatment delays. Human-centered attacks erode trust and enable broader network breaches; while systemic vulnerabilities disrupt essential services and supply chains. In this discussion it is a focal point to highlight ransomware attacks as they can halt pharmaceutical, payment or record systems. Overlapping attack surfaces highlight the need for comprehensive, layered security measures that address not only technical flaws but also human behaviors and systemic weaknesses in this uniquely sensitive environment centered around very human factors that depend on technical systems. Protecting these domains is essential for safeguarding both operational functionality and the ontological security of patients and providers. Auxiliary systems, including HVAC controls, facility access systems, supply chain logistics, and point of sale (POS) devices, are frequently overlooked in security planning despite their critical

roles in hospital operations (Millett et al, 2019; Neprash et al 2022; Al-Qarni, 2023; Li et al, 2024; Li et al, 2025; Park and Lim, 2025). For example, HVAC tampering can compromise climate-sensitive medications or equipment, while POS breaches can expose financial data and disrupt billing processes. These systems are particularly vulnerable when they are not properly segmented from core clinical networks or when security configurations lapse due to perceived low priority (Muthuppalaniappen & Stevenson, 2020).

Patients and providers need not fall to the conceit that malicious actors will not stack newly developed technologies to amplify their interference and planned disruption. For example, artificial intelligence can be used to amplify quality and quantity of attacks, to improve the odds of successful breaches and or disruption (AlZubi et al, 2021; Bonagiri et al, 2024). The discussion of AI agents and its involvement in supporting threat actors has been a point of contention (AlZubi et al, 2021; Bonagiri et al, 2024; Qiu et al, 2024; Jumaah et al, 2025). Generative AI applied to conversations and records without consent can result in a vector for exploration. Healthcare systems and insurance companies, regardless of who owns them, are at risk depending on the scale of attacks that occur at this intersection (Yeo and Banfield, 2022; Brilhante et al, 2025). Further, elements sustaining such as conceptualized in Figure 1, draw important salient points on how ontological security can be meaningfully degraded.

4. Ontological Security Concerns and Implications

The concept of ontological security refers to a state where an individual perceives their existence as authentic, intact, protected, and secure from the external factors in everyday life. In the context of healthcare, ontological security manifests as patient and provider confidence in the reliability and safety of medical care systems and environments (Griffin et al., 2023). Cyber-physical attacks on healthcare infrastructure directly undermine this fundamental sense of safety, generating profound implications for individual wellbeing and institutional trust.

Technical vulnerabilities precipitate ontological insecurity with their impact. This includes violating bodily autonomy, privacy, and personal identity. Compromised medical devices can transform therapeutic tools into sources of threat; with potential of eroding patient trust in their own bodies, treatment protocol, prescriptions, and the medical system as a whole. Biometric data theft further destabilizes identity integrity, as stolen genetic or physiological information represents an immutable personal exposure. Such violations transcend financial or privacy harms, striking at the core of an individual's perceived safety in medical care at their most vulnerable moments. Home, Holistic or Alternative Care should be discussed here as it belongs to the hemisphere of this subject matter, however, due to homeopathic belief systems it has departed from this ecosystem. While not directly intertwined with the hospital healthcare infrastructure, digital technologies are widely integrated into modern societal life. The common use of mobile applications, website archives, digital health platforms, and virtual support communities are worth noting as they may be used to inform or derive actions in members of this healthcare community. This can serve as a unique target or point of influence.

In home environments personal devices such as “smart” or integrated personal first aid devices, OTG scopes, cameras, and allocated monitors can deliver meaningful value in interaction with healthcare personnel as patients seek to gather knowledge to share or self-diagnose. However, these at-home care products with a goal for intended domestic use can differ considerably more in function and quality than hospital grade devices in terms of manufacture, security considerations, and components. While this portion of subject matter is listed as a point of reference, the topic has the depth and breadth to facilitate a separate discussion. (Dey et al, 2018; Ghirardello et al, 2018; Plappert et al, 2021; Parsons et al, 2023).

Human-centered attacks amplify relational and institutional distrust through a variety of mechanisms. Social engineering campaigns targeting clinicians corrode professional autonomy and introduce moral injury when individuals become the conduit for a breach of security. Patients subjected to coercion or manipulation experience eroded sense of self-reliance in decision making or personal authority, while broader institutional breaches foster collective uncertainty. Healthcare workers in turn can experience increased stress related to post-event scrutiny, leading to burnout and job dissatisfaction (Alanazi, 2023). To illustrate this, by examining the effects of pandemics, the misinformation distributed and the general consensus it can be observed that these effects significantly intensified (Muthuppalaniappan & Stevenson, 2020; Dixon, 2023). These periods of time as of late tend to generate an environment for accelerated digital transformation while sustaining weakened support structures or analysis.

Systemic disruptions, overall, create environmental and temporal instability that further undermine ontological security. For examination, supply chain attacks such as those targeting blood product deliveries

sabotage the predictable infrastructure necessary for continuous care (Vukotich, 2023). Ransomware-induced service delays fracture expectations of treatment, creating uncertainty about future care availability and integrity (Alanazi, 2023). Even auxiliary system compromises transform clinical environments to extensions of a landscape containing a potential threat. Factors of this scenario affect both patient care and medical supply integrity. Medical device manufacturers have responsibilities centered around the duty to warn. In the medical device industry, manufacturers are often considered to have fulfilled their obligation to warn end users (patients) by providing adequate warnings and instructions to the formally educated intermediaries (medical professionals and the adjacent) who administer these devices. The rationale utilized to validate this stance conveys that healthcare professionals are “better positioned” to understand the complexities of a manufacturer's medical devices and potential risks (Bown, 2017). The professionals slated to communicate these effectively to the patients based on individual cases are thus implied to be the "responsible" party.

Implications discussed here necessitate redefining healthcare cybersecurity beyond technical compliance. Protecting ontological security requires frameworks that address these niche threats, ensuring that encompassing defense strategies safeguard not only data and devices, but also those the system seeks to support. This approach recognizes that maintaining ontological security is essential for ensuring patient confidence and supporting healthcare providers in delivering service (Steele, 2008).

5. Cyber-Physical Risk Mitigation Considerations

Healthcare is the ultimate critical infrastructure. Securing IT systems and infrastructure within Healthcare and related Biomanufacturing, are the cornerstones of maintaining best practices and improving security posture.

Healthcare organizations face a wide range of cyber threats. These threats have the capability to disrupt patient care, steal, or alter sensitive data such as PII, PHI, financial information, or disrupt critical infrastructure that organizations need to safely operate. An example of critical infrastructure systems in organizations are in Building Management Systems (BMS) that control clean room pressures, heating and cooling systems, lighting, power and water purification systems.

Conceptually, the latest addition to the factors listed is the Internet of Biological Things (IoBT) has been discussed among scientific practitioner communities, and it is recommended that this community develops secure-by-design principles for the IoBT before the underpinning biotechnologies are incorporated in mass.

Effective Healthcare Operational Technology (OT) cybersecurity relies on defense in depth. This approach combines people, processes, and technology. Core practices include ICS-specific incident response planning, network segmentation, secure remote access, risk-based vulnerability management, and continuous monitoring. Standard IT practices also apply, such as patching, strong credentials, and embedding security into procurement. However, OT patches must be tested in isolated environments before deployment to avoid operational disruptions or malicious updates.

Effective defense of healthcare environments requires integrated mitigation frameworks that address technical, human, and systemic vulnerabilities, as discussed. These frameworks must extend beyond conventional Cybersecurity to incorporate sociotechnical principles and ontological safety objectives. These must align with Healthcare IT Security with general governance as there is work toward enhancing trust (Ewoh & Vartiainen, 2024). A multi-layered approach is essential to protect both operational functionality and the foundations of care delivery.

Technical mitigation begins with architectural resilience. Zero-trust security models, which assume no implicit trust for any user or device, provide a foundational strategy for protecting networked medical devices and electronic health records (Patel et al, 2023). Network segmentation and strict access control, isolate critical clinical systems from auxiliary infrastructure such as Building Automation Systems (BAS) e.g.; HVAC, payment systems, and adjacent support dashboards. This containment limits lateral movement by attackers, a key recommendation for protecting auxiliary systems (Muthuppalaniappen & Stevenson, 2020). Regular vulnerability assessments and patch management protocols are necessary to address weaknesses in medical device software and hospital operating systems. As this subject matter advances, AI powered solutions are becoming a more common proposition for anomaly detection or AI-driven threat hunting, with a goal to enable proactive identification of issues before they disrupt clinical operations.

OT environments have overlapping technologies with IT systems. However, the priorities and purpose of each system are very different. OT environments focus on availability and safety, where IT systems typically focus on confidentiality while addressing integrity and availability. OT systems are designed to operate for ten to

fifteen years or longer which can lead to unsupported technology without existing support, legacy systems in this case cannot easily be replaced without pairing a successor, large expense, and coordination. Developing a strong OT security program includes coordinated policies, training, and clearly defined roles. Ongoing auditing will ensure both the cybersecurity of OT systems and the physical system remains in a reliable operational state geared toward servicing patient care. Figure 2 provides examples of attacks to Healthcare OT Systems.

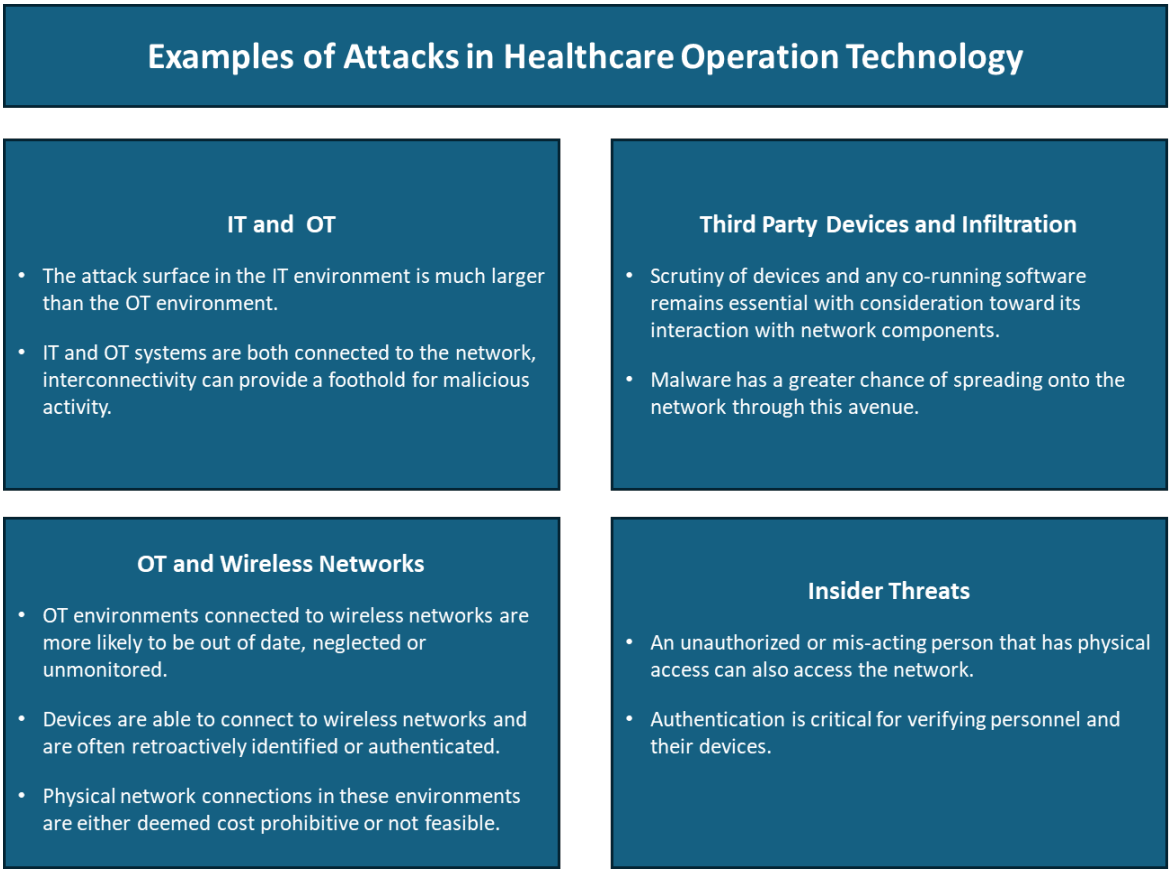


Figure 2: Examples of Attacks to Healthcare Operational Technology (OT) Systems. Attacks and misconfiguration can cause operational disruption or avenues for malicious activity

6. Discussion: Long-Term and Ethical Considerations

When we reflect on the concept of "do no harm" there exists importance in forward adjusting how we reasonably reduce harm for patients through ensuring safety. Ethics will be condensed around how urgency is a factor towards promoting better ontological security for patients and providers. Technological advancements continue to be introduced to patients, providers and hospitals and have been reactive to cybersecurity components and practices, including "Shadow IT". IT systems and components not approved officially by IT, as a supplement to approved systems define Shadow IT (Silic and Back, 2014; Raković et al, 2020; Patel et al, 2023).

There exists continuing value in assessing how actors may present unauthorized threats to existing infrastructure and how countermeasures may need to evolve over time. For example, health security research from the recent past demonstrates the ability of actors to introduce malicious hardware/software or bypass internal systems to deploy damaging or extractive action as more plausible especially, with the addition of human interaction and physical security elements (Mirsky et al, 2019; Wasserman and Wasserman, 2022; Marti, 2025).

Those masquerading as active patients, next-of-kin, personnel, or visitors can easily record conversations, as a scenario for examination, where sensitive information can be captured and utilized to manipulate individuals or systems. Actors may be able to engage with abandoned computers, as physicians and staff are distracted or not present. Consumer level and concealable tools that can clone electronic assets from a close proximity or infiltrate devices through malicious cables allow actors to gather data and execute hardware-based attacks.

Malicious software introduced into hospital workflows can be designed to rapidly learn and adapt to countermeasures to evade detection for longer, learn more about their host architecture or grow more potent. These are examples of the ability for actors to integrate into the fast paced and rapidly evolving care environment where healthcare workers practice and serve, often in cadence with one another.

Additionally, acquisitions of medical institutions by larger firms guided by an opportunity for financial gain or a business decision, versus a client focused or intentional accumulation, can pose a risk. Often the aim is to manage facilities and their operations at minimal cost. This exacerbates the vulnerability in maintaining medical infrastructure in its entirety including managing and staffing. In this environment, security is not prioritized. Access to a facility, access-controlled areas, PII or credentials can be easily acquired through abusing poorly executed practices or misinformed and undertrained personnel.

Threats like these, stand to be amplified as more technological integration continues and arguably already are. Scenarios discussed here highlight conditions of failure in ontological integrity. Reduced patient patronage as well as increased distrust of providers (and their institutions) can have a cascading effect on the degradation of applied healthcare. Solutions include what can be derived from practices that emphasize bolstering security measures and maintaining a transparent dialogue about emerging threats.

7. Limitations and Future Work

This work is a commentary, collecting considerations across a mix of works that also include conference proceedings and non-peer reviewed work. It takes on more of a conceptual nature and does not claim quantified analysis. Future work may involve further focused analytical work on ontological cyberbiosecurity perceptions focused in healthcare.

8. Conclusion

This paper explores how the increasing digitization of healthcare, from electronic health records to networked medical devices, creates a unique environment for vulnerabilities that undermine ontological security. The foundational sense of existential safety and trust for patients and providers is at the primary emphasis. Utilizing insights from prior works explored, this paper categorizes these attack surfaces into technical, human, and systemic domains. This outline serves as a canvas, detailing how compromises can lead to issues like identity fragmentation, care delays, and relational distrust. Research highlights the profound consequences of these vulnerabilities, including increased patient care avoidance and clinician burnout. The authors propose a socio-technical mitigation framework that combines zero-trust architectures, staff training, and cross-sector collaboration to reinforce ontological security. Cybersecurity considerations in the Healthcare environment are a commitment to human safety in a digitized world, aiming to provide a framework for global health institutions to protect patient safety, instil provider integrity and support institutional resilience.

Acknowledgements

Special thoughts of appreciation go out to/among staff within Hospitals, Medical Device Manufacturers and Technicians, Those of the Hacker Community Devoted to Curiosity, Improvement, and Collaboration. Additional acknowledgements go out to Ron Stephenson, Dr. Srdjan Lesaja, C.B, C.W, I.M., J.W., J.P., K.S., M.H., and others who contributed to an alternative iteration of this work.

Ethics Declaration: Ethical clearance for the research referred to in this paper was not required as no experiments were conducted.

AI Declaration: Google Docs tools were used to organize writing within this manuscript.

References

Adeoye, S.O., Lindberg, H., Bagby, B., Brown, A.M., Batarseh, F.A. and Kaufman, E.K., 2023. Cyberbiosecurity Workforce Preparation: Education at the Convergence of Cybersecurity and Biosecurity. *NACTA Journal*, 67, pp.341-351.

Alavi, A., Anvari, S., Isautier, J. M. J., & Seoudi, N. (2024). The cybersecurity landscape of telemedicine: Threats and solutions. *npj Digital Medicine*, 7(1), 44.

Aloufi, R.A.M., Alsuhaymi, N.K.N., Almutairi, A.M.M., Aljohani, M.D.D., Altarjami, R.D.D., Alharbi, A.A.S., Alrehaili, R.A.N., Alraddadi, A.E.A., Alrehaili, R.A.N., Alraddadi, S.E.A. and Alruwaythi, T.B.M., 2025. Cybersecurity and Physical Security Integration in Healthcare Institutions: Protecting Data, Infrastructure, and Human Lives. *Cultura: International Journal of Philosophy of Culture and Axiology*, 22(5s), pp.45-51.

Al-Qarni, E.A., 2023. Cybersecurity in healthcare: A review of recent attacks and mitigation strategies. *International Journal of Advanced Computer Science and Applications*, 14(5).

AlZubi, A. A., Al-Maitah, M., & Alarifi, A. (2021). Cyber-attack detection in healthcare using cyber-physical system and machine learning techniques. *Soft Computing*, 25(18), 12319-12332.

Bonagiri, K., VS, N.M., Gopalsamy, M. and SJ, S., 2024, August. AI-driven healthcare cyber-security: protecting patient data and medical devices. In: *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. IEEE, pp. 107-112.

Brilhante, M.F., Mendonça, S., Pestana, P., Rocha, M.L. and Santos, R., 2025. Economic impact of healthcare cyber risks. *Health and Technology*, 15(3), pp.635-650.

Collier, R., 2017. NHS ransomware attack spreads worldwide. *Canadian Medical Association Journal*, 189(22), pp.E786–E787. doi:10.1503/cmaj.1095434. Available at: https://pmc.ncbi.nlm.nih.gov/articles/PMC5461132/ (Accessed: [insert date]).

Dey, N., Ashour, A. S., Shi, F., Fong, S. J., & Tavares, J. M. R. (2018). Medical cyber-physical systems: A survey. *Journal of medical systems*, 42, 1-13.

Dimitrov, W. (2020). The impact of the advanced technologies over the cyber attacks surface. In: *Artificial Intelligence and Bioinspired Computational Methods: Proceedings of the 9th Computer Science On-line Conference 2020, Vol. 2 9*. Springer International Publishing, pp. 509-518.

Dixon, T.A., 2023. The bioinformational dilemma: where bioinformational diplomacy meets cyberbiosecurity. *Australian Journal of International Affairs*, 77(2), pp.169-187.

Fouad, N.S., 2024. Cyberbiosecurity in the new normal: Cyberbio risks, pre-emptive security, and the global governance of bioinformation. *European Journal of International Security*, 9(4), pp.553-573.

Ghirardello, K., Maple, C., Ng, D., & Kearney, P. (2018, March). Cyber security of smart homes: Development of a reference architecture for attack surface analysis. In: *Living in the Internet of Things: Cybersecurity of the IoT-2018*. IET, pp. 1-10.

Gilbert, S., Ricciardi, F., Mehrali, T. and Patsakis, C., 2024. Can we learn from an imagined ransomware attack on a hospital at home platform?. *NPJ Digital Medicine*, 7(1), p.65.

Greenbaum, D. ed., 2023. *Cyberbiosecurity: A new field to deal with emerging threats*. Springer Nature.

Griffin, B., Alexander, K., Palmer, X.-L. and Potter, L., 2023. Social-Engineering, Bio-economies, and Nation-State Ontological Security: A Commentary. In: *Proceedings of the 18th International Conference on Cyber Warfare and Security*, 18(1), pp.111–118. doi:10.34190/iccws.18.1.1021.

Jumaah, M., Yassin, A.A., Abduljabbar, Z.A., Jawad, M. and Nyangaresi, V.O., 2025, April. Malware Detection in a Healthcare System via Artificial Intelligence Technology: A Review. In: *Computer Science On-line Conference*. Cham: Springer Nature Switzerland, pp. 90-113.

Kubendiran, M. & Murugaiyan, A. (2015). Ontology based Access Control Model for Healthcare System in Cloud Computing. *Indian Journal of Science and Technology*, 8, 218. 10.17485/ijst/2015/v8iS9/53617.

Li, C., Wang, J., Wang, S. and Zhang, Y., 2024. A review of IoT applications in healthcare. *Neurocomputing*, 565, p.127017.

Li, S., Surineni, K. and Prabhakaran, N., 2025. Cyber-Attacks on Hospital Systems: A Narrative Review. *The American Journal of Geriatric Psychiatry: Open Science, Education, and Practice*.

Marti, C., 2025. May 30th, 2025 Competition and Cybercrime: A Theoretical Assessment and Empirical Application.

Millett, K., Dos Santos, E. and Millett, P.D., 2019. Cyber-biosecurity risk perceptions in the biotech sector. *Frontiers in bioengineering and biotechnology*, 7, p.136.

Mirsky, Y., Mahler, T., Shelef, I. and Elovici, Y., 2019. {CT-GAN}: Malicious tampering of 3d medical imagery using deep learning. In: *28th USENIX Security Symposium (USENIX Security 19)*, pp. 461-478.

Muthuppalaniappan, M., & Stevenson, K. (2020). Healthcare cyber-attacks and the COVID-19 pandemic: An urgent threat to global health. *Journal of Medical Internet Research*, 22(8), e23410.

Neprash, H.T., McGlave, C.C., Cross, D.A., Virnig, B.A., Puskarich, M.A., Huling, J.D., Rozenshtein, A.Z. and Nikpay, S.S., 2022, December. Trends in ransomware attacks on US hospitals, clinics, and other health care delivery organizations, 2016-2021. In: *JAMA Health Forum*, 3(12), pp. e224873-e224873. American Medical Association.

Park, E. and Lim, J.H., 2025. The impact of healthcare data breaches on patient hospital visit behavior. *International Journal of Research in Marketing*.

Parsons, E. K., Panaousis, E., Loukas, G., & Sakellari, G. (2023). A survey on cyber risk management for the Internet of Things. *Applied Sciences*, 13(15), 9032.

Patel, A.U., Williams, C.L., Hart, S.N., Garcia, C.A., Durant, T.J., Cornish, T.C. and McClintock, D.S., 2023. Cybersecurity and information assurance for the clinical laboratory. *The journal of applied laboratory medicine*, 8(1), pp.145-161.

Plappert, C., Zelle, D., Gadacz, H., Rieke, R., Scheuermann, D., & Krauß, C. (2021, March). Attack surface assessment for cybersecurity engineering in the automotive domain. In: *2021 29th Euromicro international conference on parallel, distributed and network-based processing (PDP)*. IEEE, pp. 266-275.

Potter, L., Ayala, O., & Palmer, X. L. (2021, February). Biocybersecurity: a converging threat as an auxiliary to war. In: *ICCWS 2021 16th international conference on cyber warfare and security*, p. 291.

Potter, L., & Palmer, X. L. (2023). Mission-Aware Differences in Cyberbiosecurity and Biocybersecurity Policies: Prevention, Detection, and Elimination. In: *Cyberbiosecurity*. Springer, Cham, pp. 37-69.

Qiu, J., Li, L., Sun, J., Wei, H., Xu, Z., Lam, K. and Yuan, W., 2025. Emerging cyber attack risks of medical ai agents. *arXiv preprint arXiv:2504.03759*.

Rajora, R., Kumar, A., Malhotra, S. and Sharma, A., 2022, December. Data security breaches and mitigating methods in the healthcare system: A review. In: *2022 International Conference on Computational Modelling, Simulation and Optimization (ICCMSO)*. IEEE, pp. 325-330.

Raković, L., Sakal, M., Matković, P. and Marić, M., 2020. Shadow IT–systematic literature review. *Information Technology and Control*, 49(1), pp.144-160.

Razaque, A., Amsaad, F., Khan, M. J., Hariri, S., Chen, S., Siting, C., & Ji, X. (2019). Survey: Cybersecurity vulnerabilities, attacks and solutions in the medical domain. *IEEE Access*, 7, 168774-168797.

Sadeghi, Z., Alizadehsani, R., Cifci, M.A., Kausar, S., Rehman, R., Mahanta, P., Bora, P.K., Almasri, A., Alkhawaldeh, R.S., Hussain, S. and Alatas, B., 2024. A review of Explainable Artificial Intelligence in healthcare. *Computers and Electrical Engineering*, 118, p.109370.

Silic, M. and Back, A., 2014. Shadow IT–A view from behind the curtain. *Computers & Security*, 45, pp.274-283.

Skorobogatjko, A., Romanovs, A., & Kunicina, N. (2014). State of the Art in the Healthcare Cyber-physical Systems. *Information Technology and Management Science*, 17(1), 126-131.

Spanakis, E. G., Bonomi, S., Sfakianakis, S., Santucci, G., Lenti, S., Sorella, M., ... & Magalini, S. (2020, July). Cyber-attacks and threats for healthcare–a multi-layer thread analysis. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, pp. 5705-5708.

Steele, B. J. (2008). *Ontological security in international relations: Self-identity and the IR state*. Routledge.

Sultanovs, E., Skorobogatjko, A., & Romanovs, A. (2016, October). Centralized healthcare cyber-physical system's architecture development. In: *2016 57th International scientific Conference on power and electrical Engineering of Riga Technical University (RTUCON)*. IEEE, pp. 1-6.

Thomson, K. L. (2021). *Cybersecurity: reducing the attack surface*. Nelson Mandela University.

Van Devender, M. and McDonald, J.T., 2023, February. A Quantitative Risk Assessment Framework for the Cybersecurity of Networked Medical Devices. In: *International Conference on Cyber Warfare and Security*, 18(1), pp. 402-411.

Verma, R. (2022). Smart city healthcare cyber physical system: characteristics, technologies and challenges. *Wireless personal communications*, 122(2), 1413-1433.

Wasserman, L. and Wasserman, Y., 2022. Hospital cybersecurity risks and gaps: Review (for the non-cyber professional). *Frontiers in digital health*, 4, p.862221.

Yeo, L.H. and Banfield, J., 2022. Human factors in electronic health records cybersecurity breach: an exploratory analysis. *Perspectives in health information management*, 19(2), p.1i.

Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M., & Alamri, A. (2015). Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Systems Journal*, 11(1), 88-95.

The APT Paradox: Sophisticated Simplicity in Nation-State Cyber Operations (2024–2025), Trends, Detection Provenance, and Practical Gaps

Raymond Andre Hagen
Norwegian University of Science and Technology (NTNU), Norway

raymohag@stud.ntnu.no

Abstract: Advanced Persistent Threats (APTs) present a paradox in cybersecurity: sophisticated state actors use both zero day exploits and old social engineering tricks, maintaining complex infrastructure while exploiting basic misconfigurations. This study analyzes 60 verified APT campaigns from January 2024 to July 2025, providing an empirical snapshot of current threat actor behaviour, targeting patterns, and detection dynamics. Using a reproducible methodology with clear inclusion criteria based on state backing, persistence, and sophistica- tion indicators, we address four research questions: which actors are active (RQ1), what sectors they target and how this varies by actor (RQ2), which initial access methods dominate (RQ3), and who detects campaigns with what implications for visibility (RQ4). All data are archived in a public repository to enable validation and extension. Our findings reveal concentration among four primary state clusters: Russia (17 campaigns), China (16), North Korea (15), and Iran (9), accounting for 95% of attributed activity. Actor sector relationships show clear patterns: Chinese actors focus on telecommunications and government networks, Russians target diplo- matic infrastructure, North Koreans emphasize financial and cryptocurrency platforms, while Iranian operations cluster around regional events. Social engineering dominates initial access (40%), followed by web/network exploitation (21.7%) and N day exploitation (13.3%), with zero days appearing in only 8.3% of campaigns, chal- lenging assumptions about APT sophistication. Critical to defensive planning, we identify systematic detection gaps from vendor centric discovery that creates predictable blind spots in regions with limited commercial se- curity deployment and sectors using legacy infrastructure. The 18 month persistence of specific actor sector relationships indicates sustained rather than episodic interest, requiring continuous defensive evolution rather than one time responses. These findings require rethinking defensive strategies from isolated organisational responses to collaborative ecosystem approaches. The paradoxical nature of APT operations, advanced yet ba- sic, strategic yet opportunistic, reflects fundamental asymmetries in cyber conflict where attackers need only single successes while defenders must maintain continuous vigilance across expanding attack surfaces. Effective defense requires not just technical controls but coordinated, cross sector frameworks based on observed rather than theoretical threat behaviours. Scope and limitations: Findings reflect publicly reported activity within January 2024–July 2025 and may under-represent restricted disclosures (e.g., Five Eyes and allied operations). We analyse observable evidence in this window, not an exhaustive census. Practical implication: Although actors often chain techniques [Cybersecurity and Infrastructure Security Agency, 2025], we treat the first successful foothold as the decision-relevant initial access because it drives earliest containment and triage; later steps refine rather than replace these priorities.

Keywords: Advanced persistent threat (APT), Cyber threat intelligence, Initial access, Threat actor, Detection provenance

1. Introduction

Advanced Persistent Threats (APTs) are marked by strategic intent, multiphase operations, and sustained access against defined targets. Typically linked to nation state operators or comparably resourced proxies, such campaigns pursue long term collection, prepositioning, or disruption. Understanding who is active, what they target, how they gain initial access, and who brings activity to light is central to research and defense. The paradoxical nature of modern APT operations, sophisticated in capability yet often simple in execution, challenges traditional threat models and defensive assumptions. From January 2024 to July 2025, publicly reported APT activity remained high while actors retooled tradecraft. Although incidents receive extensive coverage through mostly vendor driven reporting, reproducible measurement over a clearly bounded period remains scarce, creating systematic visibility gaps that shape our understanding of the threat landscape. We address this by analyzing a verified corpus of campaigns that meet the APT definition Hagen and Helkala [2024], assembled from open sources and archived in a blinded OSF repository [Blinded, 2025]. All observed campaigns in the window are archived for transparency, while primary statistics use only the analytical subset.

Our analysis is organized around four questions: who is active (RQ1), what is targeted and how targeting varies by actor (RQ2), which initial access methods and tools dominate (RQ3), and who detects and reports campaigns, and what systematic gaps this creates (RQ4). Our analysis reveals that APT actors operate across a spectrum of sophistication, that detection ecosystems create predictable blind spots, and that effective defense requires collaborative approaches that go beyond organisational boundaries.

Definition 1 (Advanced Persistent Threat (APT))**.** Following [Hagen and Helkala, 2024], we use *APT* to denote actors with (i) state backing or state level capability, (ii) sustained, campaign level operations against specific targets over time, and (iii) advanced tradecraft (e.g., bespoke malware, identity centric operations, supply chain compromise, or exploitation of public facing applications). This definition guides inclusion and coding in §3.2.

Scope & limitations (summary). Our dataset covers publicly reported APT campaigns during January 2024–July 2025 and may under-represent restricted disclosures from intelligence alliances. We analyse observable initial access and follow-on steps, not an exhaustive census of all operations. These constraints can skew prevalence estimates and regional coverage; we discuss implications and mitigation in §3.4.

2. Background and Related Work

Advanced Persistent Threats (APTs) are commonly distinguished by stealth, persistence, and strategic targeting. Unlike opportunistic or purely financially motivated intrusions, APT campaigns are typically linked to nation states or highly capable groups operating with geopolitical objectives [Chen et al., 2014, Rid, 2013, Valeriano et al., 2018]. Such operations seek long term access, sensitive data exfiltration, and, in some cases, disruption of critical functions. Two widely used lenses structure how practitioners describe and defend against APTs. The Cyber Kill Chain [Hutchins et al., 2011] models intrusion progress across phases and highlights opportunities for early detection, while MITRE ATT&CK [Strom et al., 2020] taxonomizes adversarial techniques and supports mapping between observed behaviour and defensive controls. These frameworks enable consistent language and cyber threat intelligence (CTI) sharing, but they are not a substitute for *empirical measurement* of which actors are active, which sectors are targeted, which techniques dominate, and who detects activity at a given time. Academic work has illuminated the geopolitical logic of state cyber operations. Valeriano and Maness [Valeriano et al., 2018] characterize cyber conflict as bounded by restraint and signaling, and Rid [Rid, 2013] situates operations within broader statecraft. Much of the empirical literature, however, remains retrospective and case centric, for example Stuxnet, APT1, and SolarWinds [Langner, 2011, Willett, 2021], with fewer recent, reproducible datasets that compare actors, sectors, and techniques within the same period.

Present contribution. We assemble a verified corpus of publicly reported and attributed campaigns observed between January 2024 and July 2025 and analyze the 60 records that meet our APT definition. All observed campaigns in the window are archived for transparency in a blinded OSF repository [Blinded, 2025], while borderline cases are retained for sensitivity checks but excluded from headline statistics. Each record is coded for actor attribution, targeted sectors and regions, initial access or TTPs, and detection source. Our analyses use descriptive statistics and cross tabulations aligned with four questions: who is active (RQ1), what is targeted and how targeting varies by actor (RQ2), which methods and tools dominate (RQ3), and who detects and reports campaigns (RQ4). Documentation is retained in the online archive to support review and replication. Compared to prior case oriented studies, this paper provides a recent, cross actor and cross sector snapshot of global APT activity over an 18 month window, grounded in observable evidence and designed for reproducibility.

Conventional APT wisdom and uniform-treatment assumptions. Operational guidance and standards often prescribe uniform first-line controls that fit most organisations and sectors. Agency advisories and widely used frameworks reinforce this baseline because they must be broadly applicable National Institute of Standards and Technology [2024], Cybersecurity and Infrastructure Security Agency [2023], iso [2023], European Union Agency for Cybersecurity (ENISA) [2025]. Our evidence indicates better early outcomes when priorities are calibrated to observed actor behaviour and sector exposure: where social engineering dominates, emphasise identity-centric containment and user-facing controls; where n-day reuse is common, accelerate patch triage and change management; where living-off-the-land is typical, bias monitoring toward abuse of native tools rather than malware signatures. [European Union Agency for Cybersecurity (ENISA), 2025] We therefore retain the baseline, but argue for a thin, evidence-led overlay tuned to actual APT economy-of-effort patterns observed in our window.

3. Methodology

3.1 Dataset Rationale and Scope

We assemble a verified corpus of **74** publicly reported campaigns for replication. An explicit inclusion gate yields the analytical subset of **60** campaigns that meet our APT definition, used for all statistics; the remaining **14** records are retained for sensitivity checks but excluded from headline results. Campaigns are included if they were ongoing on 1 January 2024 or initiated between 1 January 2024 and 1 July 2025. The dataset is frozen on 25 August 2025.

3.2 Data Availability

The dataset and documentation are available in a blinded OSF repository for peer review [Blinded, 2025].

Inclusion criteria and attribution handling. All observed campaigns within the window are archived for transparency [Blinded, 2025], while primary statistics use only the 60 records that satisfy our APT definition. State proxies count as APTs only when independent reporting shows tasking, command, funding, or operational control by a state service; ideological alignment alone is insufficient. We record attribution confidence to separate epistemic uncertainty from non APT activity. Borderline entries are retained for sensitivity checks but excluded from headline statistics.

Coding rules for multi-vector campaigns. We define *initial access (IA)* as the first successful foothold that enabled execution or credentialed access within the target environment. Subsequent steps in the same campaign (e.g., credential re-use, exploitation of known CVEs, lateral movement, or living-off-the-land) are recorded separately for context. Counts for step-types can therefore exceed the number of campaigns, and percentages are reported against the relevant resolved denominator (we state *N* in each caption; cf. the IA distribution figure).

Disambiguation and tie-breaks. When multiple sources describe near-simultaneous actions, IA is assigned by (i) earliest verifiable artefact time-stamp and, if tied, (ii) the analyst narrative that establishes precondition–consequence order. Ambiguous cases are marked IA-unclear and excluded from IA-percentage denominators for that break- down.

Examples. A phishing-led credential harvest followed by VPN token reuse is coded as IA=Social Engineering, with token reuse captured as a follow-on step. Phishing followed by exploitation of an already-known vulnerability (n-day) is IA=Social Engineering, while the n-day appears only in follow-on analysis rather than redefining IA.

3.3 Operationalizing the APT Definition

We use Definition 1 as both inclusion criterion and analytical frame. Each campaign is coded on four binary dimensions: **A**ctor type, **P**ersistence, **S**ophistication, **T**argeting, and included in the analytical subset if **A** = 1 and **P** + **S** + **T** ≥ 2. Observed evidence is mapped to Indicator of Compromise categories in [Hagen and Helkala, 2024] to preserve traceability from concept to data.

3.4 Data Collection and Validation

Sources are open and independently verifiable, including vendor and research threat intelligence reports, advisories, CSIRT publications, conference material, and verified news. Searches combine general engines with curated queries. AI systems are used only for lead discovery, never as direct inputs; all entries are validated against primary sources. After deduplication and alias resolution, **60** campaigns meet the APT definition and comprise the analytical subset.

3.5 Limitations

Public reporting can be incomplete or revised, potentially biasing persistence and targeting downward. Language and regional coverage can skew visibility. We mitigate through conservative thresholds, explicit attribution confidence recording, and release of the blinded archive for review and replication [Blinded, 2025].Results should be read as a documented snapshot of observable activity in the window, not an exhaustive census.

4. Results

This section answers RQ1–RQ4 using the analytical subset. Borderline records are retained for transparency in the blinded archive but excluded from quantitative statistics [Blinded, 2025]. Unless stated otherwise, percentages use the 60 denominator.

4.1 Geopolitical Attribution of APT Campaigns (RQ1)

National attribution follows the coding in §3.2 and the curated dataset [Blinded, 2025]. Table 1 reports the counts. Minor or contested attributions are merged into *Unknown/eCrime* per the study taxonomy.

Table 1: Campaigns by attributed country or group and primary strategic focus. Minor attributions, for example India and South Korea, are merged into Unknown/eCrime

Country/Group	Number of Campaigns	Primary Target Focus (examples)
Russia	17	Ukraine and NATO/EU, diplomatic corps and webmail platforms, critical infrastructure and energy, destructive tooling and wipers
China	16	Government networks, telecom and network edge devices, Southeast Asia and Taiwan, EU governments, supply chain and identity abuse, botnets on end of life devices
North Korea	15	Cryptocurrency and financial sector, South Korean enterprises, software supply chains and developer ecosystems, defence and nuclear, technology companies
Iran	9	Israel and the broader Middle East, satellite and communications, oil & gas, aerospace and defence, spear phishing and impersonation campaigns
Unknown/eCrime	3	Enterprise intrusion and monetisation (for example FIN7, SCATTERED SPIDER, OldGremlin), cloud, SaaS, and retail; regional outliers

4.2 Actor Sector Relationships (RQ2)

Actor sector variation concentrates in a few pairings. Figure 1 summarizes the eight core sectors used for cross actor comparisons. Additional sectors and long tail categories are available in the dataset [Blinded, 2025].

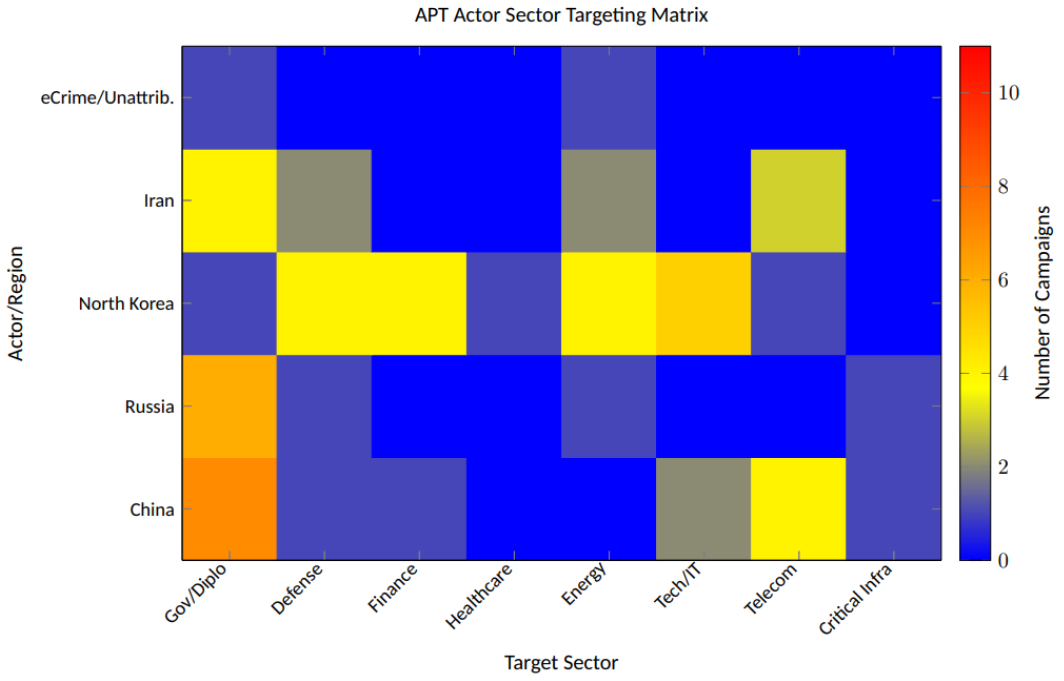


Figure 1: Actor family by sector across confirmed campaigns (N = 60; window=January 2024–July 2025). Cell counts reflect observed campaign–target pairs; totals can exceed N where a campaign targeted multiple sectors or regions

4.3 Initial Access Techniques (RQ3)

Initial access is led by a small set of vectors. Figure 2 shows campaign counts per vector.

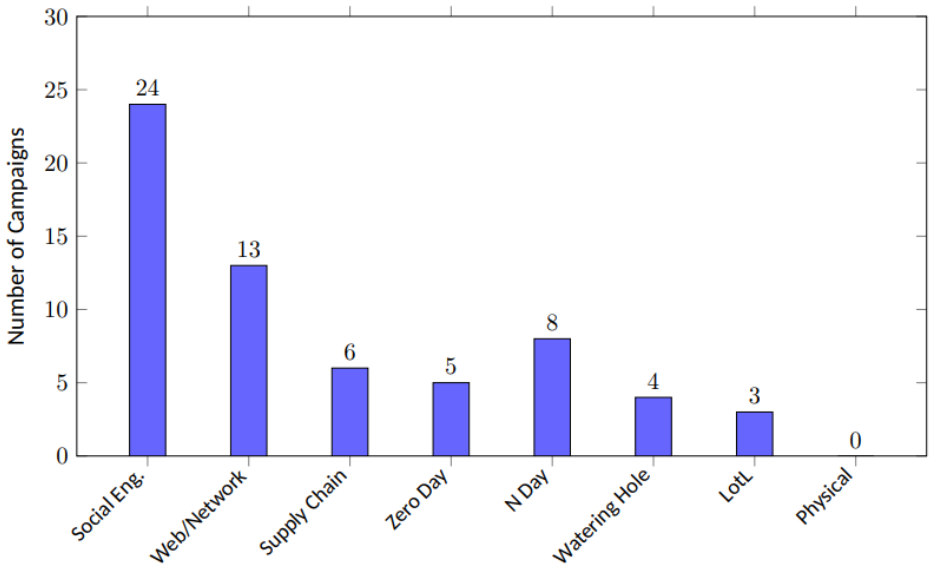


Figure 2: Initial access distribution across confirmed campaigns (N = 60; window=January 2024–July 2025). Social engineering dominates, while n-day reuse and zero-day exploitation appear less frequently; IA is defined as the first successful foothold, with follow-on steps recorded separately

Sequence-aware interpretation. Actors frequently chain multiple steps after the first foothold (e.g., credential re-use, lateral movement, or exploitation of already-known vulnerabilities). We therefore treat *initial access* as the first successful foothold that triggers the earliest containment and triage decisions, while recording subsequent steps separately for context (cf. our IA distribution figure and campaign tables). This keeps the measure decision-relevant for defenders, without denying that later steps often escalate impact.

4.4 Detection Provenance (RQ4)

Most campaigns enter the public record via vendor threat reporting, followed by coordinated advisories and CSIRTs; tech vendors contribute a smaller share. Coattributions occur but do not change the majority pattern. organisation level tallies and copublication details are available in the curated archive [Blinded, 2025].

5. Discussion

Table 2: Key metrics from the analytical subset

Top actors (counts)	Russia (17), China (16), North Korea (15), Iran (9), Unknown/eCrime (3).
Top sectors (counts)	Government/Diplomatic (19); Defense/Aerospace (8), Telecommunications (8), Energy/Nuclear (8).
Initial access (counts)	Social engineering (24), Web/Network (13), N day (8), Supply chain (6), Zero day (5).

We interpret the analytical subset in light of RQ1–RQ4. .

5.1 RQ1: Geopolitical Actor Trends

Four primary state clusters, Russia (17), China (16), North Korea (15), and Iran (9), represent 95% of attributed campaigns, revealing shifts in operational tempo, targeting discipline, and collaborative behaviours. Russia's 17 campaigns reflect sustained operational tempo despite sanctions and exposed tooling. Persistence suggests acceptance of attribution costs or confidence that adaptation outpaces defensive responses. Distribution across APT28, APT29, and Sandworm maintains GRU and SVR separation with limited tactical convergence despite shared objectives. This compartmentalization provides resilience against comprehensive takedowns. China's 16 campaigns demonstrate industrial scale operations with uniform tradecraft, suggesting centralized coordination despite distributed execution. Consistent targeting of edge devices and identity systems indicates systematic intelligence requirements from central planning rather than opportunistic targeting. Absence of financially motivated operations reinforces purely strategic objectives. North Korea's 15 campaigns reveal operational maturation, with Lazarus sustaining concurrent campaigns across sectors while maintaining operational security. Documented

handoffs between access teams and ransomware operators suggest professionalized criminal partnerships blurring state and criminal boundaries. This hybrid model may emerge for sanctions constrained actors. Iran's 9 campaigns show increased sophistication in social engineering and targeting precision. Temporal clustering around regional events suggests specific collection requirements rather than persistent broad spectrum collection. Lower volume reflects capability constraints, selective targeting, or improved operational security. Minimal other nation state representation (3 Unknown/eCrime campaigns) suggests attribution challenges for emerging actors, reduced activity, or migration to gray zone operations below APT threshold. Concentration among established actors indicates high barriers to sustained APT operations, requiring technical capability, operational infrastructure, targeting intelligence, and sustained resourcing few actors can maintain.

5.2 RQ2: Sectoral Targeting Patterns

Government and diplomatic targets lead observed campaigns (19 instances), followed by defense, telecommunications, and energy sectors (8 each). The actor sector matrix exposes distinct strategic logics. China's concentration on telecommunications (4 campaigns) and government networks (7 campaigns) aligns with technology acquisition priorities and prepositioning for regional contingencies. Emphasis on network edge devices suggests systematic mapping of global routing infrastructure for persistent access enabling both peacetime collection and wartime disruption. North Korea's diverse targeting spans technology (5), defense (4), finance (4), and energy (4), reflecting dual pressures: sanctions driven revenue generation through cryptocurrency theft and strategic intelligence supporting nuclear and missile programs. Equal distribution across sectors suggests coordinated tasking across operational units with distinct missions, marking maturation from earlier predominantly financial opera- tions. Russia's government and diplomatic focus (6 campaigns) intensified following geopolitical tensions, shift- ing from broad spectrum targeting pre-2024 to concentrated NATO decision making infrastructure collection. Absence from finance and healthcare suggests improved operational security, strategic reorientation toward mil- itary intelligence priorities, or displacement of criminal activity to nonstate actors. Emphasis on destructive capa- bilities reinforces strategic rather than economic objectives. Iran's telecommunications emphasis (3 campaigns) correlates with regional surveillance objectives and bypassing international communications monitoring. Gov- ernment and diplomatic targeting (4 campaigns) shows temporal clustering around regional events, suggesting responsive rather than persistent collection. This reactive pattern differs from sustained Chinese and Russian approaches, potentially reflecting resource constraints or limited strategic objectives. These patterns indicate sector specific defensive investments must account for concentration and actor specific motivations. Identity and privilege management prove critical for telecommunications and energy facing Chinese and Russian inter- est. Transaction monitoring addresses DPRK financial targeting. Diplomatic communication hardening counters Russian and Iranian government focused operations. Recurrence of sector actor pairings across the 18 month window suggests persistent rather than episodic interest, requiring sustained defensive postures.

5.3 RQ3: Initial Access Emphasis

Why initial access still matters. Even when campaigns involve several chained techniques, the first foothold sets the time-critical actions: identity containment, scoping, and patch triage Mandiant [2025], CrowdStrike, Inc. [2025]. Later stages refine rather than replace these priorities, so reporting the initial vector remains a useful, decision-oriented signal for defenders, whereas later steps describe escalation paths.

Social engineering's dominance (24 campaigns, 40%) over technical exploitation reveals a shift in APT at- tack economics. Combined with web/network exploitation (13 campaigns) and N day usage (8 campaigns), these account for 75% of observed initial access, suggesting most APT objectives are achievable without zero day capabilities. Social engineering transcends simple phishing, encompassing sophisticated impersonation chains, compromise of trusted platforms, and exploitation of professional networking sites. Consistency across all major actors indicates convergent evolution toward exploiting human trust rather than technical vulnerabilities. This pattern is pronounced in diplomatic and government sectors, where spear phishing leverages geopolitical events to achieve compromise. Sustained use suggests human layer defenses lag behind technical controls. Web and network exploitation (13 campaigns) targets internet facing applications and VPN endpoints, emphasizing recently disclosed enterprise vulnerabilities. Exploitation within days of disclosure indicates APT actors maintain ready weaponization capability. This rapid operationalization challenges traditional patch windows and suggests the zero day versus N day distinction matters less than time to patch metrics. Low zero day usage (5 campaigns, 8.3%) contradicts assumptions about APT sophistication. Zero days appear reserved for high value targets where social engineering is infeasible or persistent access justifies operational cost. Selective deployment suggests rational resource allocation rather than capability constraints. Supply chain compromise (6 campaigns, 10%) enables

single intrusions to cascade across victims. Clustering among Chinese and Russian actors suggests this requires sustained infrastructure and victim research smaller actors cannot maintain. Targeting software pipelines and managed service providers indicates strategic multiplier selection. Limited watering hole (4 campaigns) and living off the land techniques (3 campaigns) as initial access may reflect detection bias, as these methods are harder to attribute. Their presence in sophisticated Chinese and Russian operations suggests these techniques require extensive reconnaissance and operational patience aligned with long term collection rather than immediate exploitation.

5.4 RQ4: Detection Provenance

Campaign discovery concentration through vendor threat reporting, followed by coordinated advisories and CSIRTs, reveals critical dependencies and gaps in the global detection ecosystem. Vendor dominance raises questions about visibility bias. Commercial security vendors' detection is limited to their deployment footprint, creating blind spots in regions with limited commercial tool adoption, sectors with legacy infrastructure, and organisations using open source or indigenous solutions. Geographical distribution of major vendors in North America, Western Europe, and developed Asian markets may result in underreporting of campaigns targeting Africa, Latin America, and Central Asia. This detection inequality could enable persistent adversary presence in undermonitored regions serving as staging grounds for broader campaigns. Coordinated advisories and CSIRTs' secondary role suggests na- tional and sector specific detection capabilities rarely achieve first detection of novel campaigns. This lag reflects resource constraints, limited cross organisational visibility, or defensive rather than hunting oriented missions. Few government first detections involved campaigns directly targeting government networks, suggesting victim proximity remains crucial for discovery. Minimal contribution from technology vendors indicates persistent chal- lenges in building security observability into products. Despite repeated targeting of edge devices, productivity suites, and development tools, these vendors rarely achieve first detection of abuse. Product telemetry remains optimized for operational rather than security purposes, missing early detection opportunities at platform level. Detection to disclosure timing varies by source. Vendor led discoveries typically reach public disclosure within 30-90 days, balancing responsible disclosure with commercial imperatives. Government led discoveries experi- ence longer delays, suggesting extended remediation or intelligence gain/loss calculations delaying attribution. These differences affect defensive response to emerging threats, with vendor driven disclosure providing more actionable intelligence in tactical timeframes. Absence of victim self reporting suggests organisations lack detec- tion capabilities or reputational and regulatory concerns discourage disclosure. This silence limits understanding of dwell time, impact, and defensive effectiveness, creating an incomplete threat picture that may overemphasize detected campaigns while missing persistent, undetected intrusions.

Provenance caveats. First-sighting claims are not uniformly distributed across sensors or regions; they depend on product footprint, customer mix, and disclosure practices. As a result, "who saw it first" is partly a function of market presence rather than global prevalence. To reduce this bias, we triangulate across multiple report types (vendor advisories, CERT bulletins, and community write-ups) and time-stamp the earliest public disclosure per campaign, while keeping later confirmations for context.

5.5 Synthesis: Cross-Cutting Implications for Attribution and Defence

Where conventional wisdom comes from, and where it fails. Much of the field's narrative is produced by vendors and agencies with uneven telemetry and staggered disclosure timelines. This can over-amplify techniques that their sensors capture well and understate activity outside their footprint. Baseline guidance that treats threats uniformly is useful, but our evidence indicates better outcomes when priorities are actor- and sector- specific: identity hardening where social engineering dominates, accelerated patch triage where nday reuse is common, and targeted monitoring where livingofftheland is typical National Cyber Security Centre [2024], Cy- bersecurity and Infrastructure Security Agency [2025].

Integrated analysis of actor patterns, targeting preferences, access methods, and detection dynamics reveals systemic implications requiring reconsideration of attribution methodologies and defensive architectures.

Attribution confidence emerges as hierarchical. National clustering provides strongest signal, reinforced by predictable targeting patterns and tooling preferences within geopolitical groups. Group specific tooling chains like MISTPEN for Lazarus or custom loaders for APT29 provide secondary confirmation when combined with infrastructure overlaps. However, widespread commodity framework adoption across actors, Cobalt Strike appearing in all major nation state campaigns, diminishes attribution value. This hierarchy suggests attribution requires differential weighting: infrastructure patterns and bespoke malware carry substantial weight, commodity tools

provide minimal value, and targeting patterns serve as corroborating rather than primary evidence. organisations must rapidly assess whether activity represents targeted state interest or opportunistic criminal activity.

Defensive architectures must evolve beyond uniform controls to account for pronounced actor sector convergence. Chinese telecommunications concentration, Russian government focus, and North Korean financial emphasis enable risk calibrated implementation. While identity management emerges as universal given social engineering dominance, implementation must reflect threat specific adaptations. Telecommunications facing Chinese interest require supply chain verification and hardware attestation for edge devices. Financial institutions under North Korean targeting need transaction anomaly detection emphasizing cryptocurrency conversion and cross border transfers. Government entities confronting Russians must implement diplomatic communication isolation with air gapped authentication resistant to destructive attacks. Sector specific adaptations grounded in observed preferences yield superior outcomes compared to generic frameworks.

Vendor centric detection creates systematic gaps sophisticated actors exploit for persistent access. organisations in regions with limited commercial deployment, sectors using legacy infrastructure, or environments where living off the land predominates must assume APTs operate below detection thresholds. This necessitates shifting from detection centric models assuming timely discovery to resilience architectures assuming persistent presence. Such architectures prioritise segmentation limiting lateral movement, immutable backups for restoration, and deception technologies increasing adversary costs. Hunt operations should focus on limited vendor visibility areas: operational technology networks, non Windows environments, and legacy systems where evidence suggests persistent footholds.

Temporal dynamics require defensive strategies operating across multiple horizons simultaneously. Rapid vulnerability weaponization with N day exploitation within days demands immediate patch protocols compressing testing windows. Sustained social engineering effectiveness indicates human defenses require continuous reinforcement. The 18 month persistence of actor sector relationships demonstrates sustained strategic interest requiring continuous evolution. This complexity requires programs maintaining simultaneous focus on immediate vulnerability response, sustained awareness training, and long term transformation, with resource allocation reflecting relative risk.

These findings challenge conventional wisdom treating threats uniformly and assuming technical controls adequately defend against nation states. Evidence supports differentiated models where attribution drives response, sector determines prioritization, visibility gaps guide hunting, and temporal dynamics shape planning. organisations facing APT threats must develop adaptive programs reflecting sector, geography, and specific adversaries likely targeting their assets, with investments calibrated to observed rather than theoretical behaviours.

5.6 Network Dynamics and Collaboration Patterns

Network analysis reveals distinct operational patterns across actor groups. North Korean operations demonstrate cohesion with limited cross actor linkages. Two documented bridges underpin the DPRK subgraph: Andariel's initial access handoff to Play ransomware, and the Lazarus UNC2970 linkage via MISTPEN chain. Beyond these, overlaps remain at tradecraft level rather than direct infrastructure or tooling sharing. Russian operations exhibit compartmentalization across intelligence services. APT28, APT29, and Sandworm show limited evidence of direct implant or command and control infrastructure sharing, with similarities confined to thematic lures and targeting patterns. This separation aligns with documented interservice boundaries between GRU and SVR, suggesting organisational silos persist even when objectives converge. Chinese activity manifests through parallel operational cells emphasizing telecommunications infrastructure and identity systems. Clusters concentrate on network edge devices and supply chain compromise, developing bespoke malware within distinct cells while demonstrating limited intercluster tool reuse in public campaigns. This pattern suggests coordinated strategic direction with distributed tactical execution. Widespread commodity framework presence across actors underscores an attribution principle: shared tooling does not constitute collaboration evidence. Frameworks like Cobalt Strike appear across multiple nation state operations and must be excluded as adjacency evidence. Only group specific tooling chains with documented operational linkages, such as MISTPEN between Lazarus and UNC2970, provide valid actor relationship evidence.

Decision-oriented implications. Given the IA distribution in Fig. 2, prioritise identity-centric containment over rare zero-day scenarios. Because re-use of known vulnerabilities materially exceeds zero-day use in our window, patch triage against the currently exploited set should outrank generic "patch everything" drives. Use the actor×sector heatmap (Fig. 1) to add a thin, sector-specific overlay to the baseline rather than treating all threats

uniformly. Finally, since first-sighting is partly a function of telemetry footprint, require multi-source corroboration before rebalancing investments on the basis of a single “who saw it first” claim.

6. Conclusion

We analyzed 60 verified APT campaigns from January 2024 to July 2025, revealing a paradoxical threat landscape. APTs deploy zero day exploits while relying on decades old social engineering, maintain complex infrastructure while exploiting basic configuration errors, and pursue strategic nation state objectives while adapting opportunistically. This duality challenges conventional models positioning APTs as uniformly sophisticated, instead revealing actors who select minimum viable techniques to achieve objectives. Activity concentration among four nation state clusters, Russia (17), China (16), North Korea (15), and Iran (9), demonstrates tight coupling between geopolitical tensions and cyber operations. Within this state directed framework, we observe pragmatic adaptations driven by resource constraints, operational security, and target availability. North Korea’s dual focus on revenue and strategic intelligence exemplifies tension between state direction and operational autonomy. Documented handoffs between nation state operators and criminal ransomware groups blur traditional boundaries, suggesting APTs operate across a spectrum rather than within rigid frameworks. For defenders, this complexity strains organisational capabilities. Campaigns exploited vulnerabilities from firmware implants to social manipulation, demanding expertise across widening technology arrays. organisations in telecommunications and government face compound challenges where measures optimized for one adversary prove inadequate against another. Rapid vulnerability weaponization alongside persistent legacy techniques requires programs operating across multiple horizons, maintaining immediate response and long term transformation simultaneously. Systematic visibility gaps from detection provenance analysis reveal fragmented, incomplete defensive capabilities. Detection concentration in commercial vendors creates blind spots sophisticated actors exploit, while limited technology vendor contribution suggests persistent disconnects between product development and security requirements. Gaps extend beyond technical capabilities to expertise, where the skill gap widens despite increased investment. Evidence suggests individual organisational responses cannot adequately address APT threats leveraging global infrastructure and cross border flexibility. These findings compel reconsideration from isolated organisational responses to collaborative ecosystem approaches. Cross sector sharing must evolve beyond indicators to include behavioural patterns, attribution methodologies, and defensive innovations raising collective resilience. International cooperation becomes essential for real time defensive coordination matching transnational adversary tempo. Actor sector persistence indicates targeted sectors share defensive challenges benefiting from collaboratively developed rather than duplicated frameworks. The paradoxical APT nature, advanced yet basic, strategic yet opportunistic, persistent yet adaptive, reflects fundamental asymmetry where attackers need single successes while defenders maintain continuous vigilance across expanding surfaces. This asymmetry requires systemic approaches acknowledging geopolitical drivers and operational realities of defending interconnected systems. As APTs evolve tradecraft responding to improvements and attribution, collaborative, adaptive, strategically informed approaches become essential for organisational and national security.

7. Future Work

Extending the observation window to multiple years would enable longitudinal analysis of campaign evolution, ac- tor adaptation, and correlation with geopolitical events. A five year dataset could capture cyclical patterns tied to political transitions or economic cycles invisible in the current 18 month snapshot. Geographic parameters including attack infrastructure geolocation, operational hours, and network routing paths would enable spatial temporal pattern recognition revealing operational security practices. Technical telemetry at packet and behavioural levels would support machine learning approaches for attribution. Integration of sanctions regimes, diplomatic ten- sions, and military readiness indicators could reveal predictive relationships between real world events and cyber operations. The actor relationship edge list provides foundation for network analysis the current dataset cannot support. With expanded coverage, graph neural networks trained on actor infrastructure target relationships could reveal latent patterns. Time series analysis might identify critical transitions preceding campaign intensifi- cation or collaboration. Predictive modeling could transform defense from reactive to anticipatory. Classification models could assess attribution probability given partial observables. Survival analysis could model time to de- tection based on sector, actor, and technique combinations. Agent based simulations with observed behaviours could explore intervention strategies. Incorporating non English sources and regional CERT partnerships would reduce Western bias and surface invisible campaigns. Standardized coding schemas with inter rater reliability protocols would enable distributed collection while maintaining quality. The objective extends beyond analysis to operational impact. A continuously updated dataset with rich attributes would enable evidence based strate- gies, quantitative risk

assessment, and optimal resource allocation. Transition from static analysis to dynamic prediction would fundamentally alter how organisations respond to nation state threats, moving from anecdotal intelligence to data driven strategies grounded in empirical adversary behaviour.

Declaration of generative AI and AI-assisted technologies in the writing process: The author used AI assisted tools to support language editing and formatting: (i) improving clarity, grammar, and idiom; (ii) reducing dyslexia related spelling/ordering errors; and (iii) troubleshooting LATEX minutiae (bibliogra- phy keys, table widths, listing frames). All AI suggested edits were reviewed and approved by the author. Any remaining typos are proudly human.

Ethics declaration: This research uses exclusively publicly available open source data, requiring no ethical clearance. The dataset is anonymized for review to protect author identities, consistent with double blind review protocols.

References

Blake Strom, Andy Applebaum, Douglas Miller, Kathryn Nickels, Adam Pennington, and Cody Thomas. Mitre att&ck®: Design and philosophy. Technical Report PRS 19-01075-28, The MITRE Corporation, March 2020. URL https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf.

Brandon Valeriano, Benjamin Jensen, and Ryan C. Maness. *Cyber Strategy: The Evolving Character of Power and Coercion*. Oxford University Press, 2018. ISBN 978-0-19-061809-4.

CrowdStrike, Inc. 2025 global threat report, February 2025. URL https://www.crowdstrike.com/en-us/ global-threat-report/.

Cybersecurity and Infrastructure Security Agency. Cross-sector cybersecurity performance goals (cpgs), Octo- ber 2023. URL https://www.cisa.gov/cross-sector-cybersecurity-performance-goals. Baseline practices applicable across critical infrastructure.

Cybersecurity and Infrastructure Security Agency. Known exploited vulnerabilities (kev) catalog, 2025. URL https://www.cisa.gov/known-exploited-vulnerabilities-catalog. Live catalogue, updated con- tinuously.

Dataset Blinded. Apt dataset 2024–2025, 2025. URL https://osf.io/bhv53/?view_only=.

Eric M. Hutchins, Michael J. Cloppert, and Rohan M. Amin. Intelligence-driven computer network defense in- formed by analysis of adversary campaigns and intrusion kill chains. In *Proceedings of the 6th International Conference on Information Warfare and Security (ICIW)*, pages 113–125. Academic Conferences International Limited, 2011.

European Union Agency for Cybersecurity (ENISA). Enisa threat landscape 2025, October 2025. URL https://www.enisa.europa.eu/sites/default/files/2025-10/ENISA%20Threat%20Landscape%202025.pdf. TLP:CLEAR.

Information technology — security techniques — information security incident management — part 1: Principles of incident management, 2023. URL https://www.iso.org/standard/78973.html.

Mandiant. M-trends 2025 report, April 2025. URL https://services.google.com/fh/files/misc/m-trends-2025-en.pdf.

Marcus Willett. Lessons of the solarwinds hack. *Survival*, 63(2):7–26, 2021. doi: 10.1080/00396338.2021. 1906001. URL https://www.tandfonline.com/doi/full/10.1080/00396338.2021.1906001.

National Cyber Security Centre. Recommended types of mfa, September 2024. URL https://www.ncsc.gov.uk/collection/mfa-for-your-corporate-online-services/recommended-types-of-mfa. Guidance for corporate online services.

National Institute of Standards and Technology. The nist cybersecurity framework (csf) 2.0. NIST Cybersecurity White Paper NIST CSWP 29, NIST, February 2024. URL https://doi.org/10.6028/NIST.CSWP.29.

Peter Chen, Lieven Desmet, and Christophe Huygens. A study of advanced persistent threats. *IFIP In- ternational Conference on Communications and Multimedia Security*, pages 63–72, 2014. doi: 10.1007/ 978-3-662-44885-4 5.

Ralph Langner. Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, 9(3):49–51, 2011. doi: 10.1109/MSP.2011.67.

Raymond Andre´ Hagen and Kirsi Helkala. The complexity of contemporary indicators of compromise. In *Pro- ceedings of the 23rd European Conference on Cyber Warfare and Security (ECCWS 2024)*, pages 693–703, Reading, UK, June 2024. Academic Conferences International Limited. doi: 10.34190/eccws.23.1.2149. URL https://papers.academic-conferences.org/index.php/eccws/article/view/2149.

Thomas Rid. *Cyber War Will Not Take Place*. Oxford University Press, 2013. ISBN 978-0-19-933063-8.

Quantifying the Economic Impact of Ransomware: Cyber Risk Modeling with Gamma Regression

Li Huang and Kimberly Cornell
University at Albany, USA

lhuang9@albany.edu
kacornell@albany.edu

Abstract: Ransomware has evolved into one of the most disruptive forms of cybercrime, resulting in severe financial and operational losses for organizations across various sectors. Despite its increasing prevalence, quantitative models for evaluating the drivers of ransomware losses remain underdeveloped, limiting both academic insight and practical risk management. This study addresses this gap by developing and empirically validating a statistical framework for quantifying the financial impact of ransomware incidents. Drawing on ransomware cases extracted from the Advisen Cyber Loss Database, we employ a generalized linear model (GLM) with Gamma regression and log link to estimate how socio-technical factors shape loss severity. Our analysis examines four categories of predictors: (1) Technology, operationalized via database server involvement in incidents; (2) Preparedness, captured through insurance coverage ratios; (3) Settlement Length, reflecting negotiation and resolution timelines; and (4) Multi-entity Connections, representing the number of affected organizations in an incident. The results indicate that each factor influences the anticipated magnitude of losses. Insufficient preparedness is correlated with greater financial damages, whereas incidents involving database servers and prolonged settlement periods result in disproportionately substantial losses. Moreover, multi-entity connections exacerbate losses due to cascading effects across organizational networks. To assess the robustness of the model, bootstrapping techniques are employed, confirming the stability of the coefficient estimates and underscoring the model's reliability under resampling. By providing empirical evidence of the drivers of ransomware loss severity, this study contributes to both academic research and practical cybersecurity governance. For scholars, it demonstrates the utility of Gamma regression in modeling highly skewed cyber loss distributions. For practitioners, it highlights quantifiable indicators that can inform cybersecurity investment and organizational preparedness strategies. More broadly, the findings highlight the importance of interdisciplinary approaches that integrate cybersecurity management with socio-technical dimensions of cyber risk. This work lays a foundation for future studies that extend to newer datasets and explore AI-enhanced risk prediction, thereby advancing both theoretical understanding and applied resilience against ransomware in the evolving cyber threat landscape.

Keywords: Ransomware attacks, Cyber risk management, Quantitative modelling, Organizational resilience

1. Introduction

Ransomware has become a targeted attack, primarily aimed at maximizing monetary gains (August, et al., 2019). However, modern ransomware operations have shifted toward a targeted attack, where attackers carefully select victims based on system vulnerability, data sensitivity, and willingness to pay a ransom (OZ, et al., 2022). This shift is driven by profit-maximization, as attackers seek to extract the highest possible payouts (Samonas, et al., 2020). Since ransomware attacks are deliberately targeted rather than purely opportunistic, organizations must assess their specific risk profiles rather than relying solely on generic malware defenses. In evaluating ransomware risks, non-technical factors, such as industry characteristics, cybersecurity preparedness, and overall vulnerability, play a crucial role in determining the impact of ransomware attacks. Understanding these factors enables more accurate risk assessments and the implementation of proactive defense.

Non-technical factors, such as industry sector, shape how an organization responds to an attack and influence the overall expected losses. For instance, organizations in critical sectors (e.g., healthcare, finance, and law enforcement) prioritize operational continuity and the security of sensitive information (Dawson, et al., 2021; Tsen, et al., 2022). By incorporating these sociotechnical factors, cybersecurity researchers can better assess the full impact of ransomware incidents and implement more effective mitigation measures (Samonas, et al., 2020). Yet, to validate the influence of these factors, reliable empirical data is essential.

However, one persistent challenge in ransomware research is the limited availability of high-quality empirical data. Many studies rely on simulated data, limited case studies, or proprietary industry reports that lack transparency or reproducibility (Anderson, et al., 2013). Due to privacy concerns and reputational risks, organizations underreport ransomware incidents, which results in under-sampled and biased datasets (Dekker & Karsberg, n.d.; Rassam, et al., 2017). Without comprehensive empirical data, it is difficult to validate theoretical models or accurately quantify economic impact. As a result, despite growing scholarly and industrial attention, limited research has quantitatively examined how socio-technical factors determine the financial impact of ransomware attacks. Prior studies have mainly focused on technical detection and response

mechanisms, leaving a gap in data-driven understanding of the economic consequences and risk drivers of ransomware.

To address this gap, we develop a quantitative framework that combines insights from cyber risk management and actuarial modeling. Drawing on the Advisen Cyber Loss Database (Advisen Insurance Intelligence, 2015) that integrates comprehensive data on cyber incidents resulting in substantial monetary losses, we empirically estimate ransomware loss severity using a Gamma regression model validated through bootstrapping. This approach advances current discourse by providing statistically grounded evidence on the drivers of ransomware losses and by identifying measurable indicators that can inform both academic research and practical cybersecurity management. Our focus encompasses the following: initially, we analyze the factors that influence the impact of ransomware risks, particularly those analogous to the determinants used in estimating the impact of data breaches. These factors include organizational structures, technological components, and vulnerabilities. Subsequently, we assess the monetary implications associated with ransomware attacks. According to McIntosh et al., ransomware losses are affected by attack duration, vulnerabilities, technological aspects, and connectivity (McIntosh, et al., 2021).

The remainder of this paper is organized as follows. Section 2 reviews the existing literature on ransomware and socio-technical determinants of cyber risk. Section 3 outlines the quantitative research framework, data description, and variable construction. Section 4 presents the empirical model, estimation procedures, and validation strategy. Section 5 analyzes the results and discusses implications for cybersecurity management. Finally, Section 6 concludes with key findings, limitations, and directions for future research.

2. Literature Review

Ransomware has evolved from opportunistic attacks to sophisticated, profit-driven operations that exploit both technical vulnerabilities and organizational weaknesses (August, et al., 2019; OZ, et al., 2022). This shift has intensified the financial burden on organizations and has drawn increasing scholarly attention. However, existing ransomware studies remain fragmented across multiple disciplines, including computer science, information systems, economics, and organizational studies. This disciplinary fragmentation often leads to isolated approaches to understanding and addressing the fast-evolving cyber threat (Martin & Collier, 2020; Neoaz, 2024). Most prior studies examine technical detection or incident response rather than the socioeconomic determinants of ransomware impact (Samonas, et al., 2020; Mukhopadhyay & Jain, 2024). Yet a comprehensive knowledge of ransomware requires an integrated perspective that considers technology, organizational behavior, and economic incentives. Despite a recognized need for interdisciplinary approaches, few studies have attempted to integrate socio-technical dimensions while quantifying ransomware risks (Mukhopadhyay & Jain, 2024).

In addition, quantitative analysis of ransomware loss severity remains limited. Although conceptual frameworks exist for valuing cyber risk, empirical validation with incident-level data is scarce (Anderson, et al., 2013). Moreover, recent research emphasizes that ransomware losses follow heavy-tailed, highly skewed distributions, requiring statistical methods beyond classical ordinary-least-squares assumptions (Mukhopadhyay, et al., 2019; Dutta & Perry, 2007; Edwards, et al., 2016). Gamma model best fits for positive, highly skewed data and is widely used in real-world applications including cyber-loss modelling (Ben-Assuli & Padman, 2020). However, few studies employ Gamma regression to capture cyber loss patterns. Addressing this methodological gap, we integrate socio-technical variables into a Gamma-based framework to estimate how the influencing factors discussed in the existing literature determine the impact of ransomware incidents.

Prior work identifies several socio-technical determinants of ransomware impact, summarized in Table 1.

Table 1: Synthesis for Socio-technical Determinants

Socio-technical Factor	Relevant Literature
Technology	First, the extent of technology usage in operational and business services significantly influences ransomware risks. For example, studies indicate that an increase in digital records correlates with a higher risk of ransomware attacks (Kim & Kwon, 2019). However, the impact of such attacks varies based on the specific IT components compromised. For instance, encrypting an internal system may result in lower losses compared to a server breach, as servers often contain more sensitive and critical information (Hoferek & Wilson, 2007). Furthermore, the number of digital assets influences potential ransomware risks. The more devices, applications, servers, and network components a company has, the greater the attack surface, and potentially more vulnerabilities to ransomware threats.

Socio-technical Factor	Relevant Literature
Vulnerability	An organization's IT environment plays a crucial role in assessing ransomware risks. No organization's cybersecurity is entirely immune to vulnerabilities that attackers can exploit (Gordon & Loeb, 2003). These vulnerabilities may exist in hardware, software, external connections, employee behavior, etc. Frameworks such as those provided by NIST offer guidance to enhance an organization's defenses against ransomware threats. The level of preparedness is directly related to the extent of these vulnerabilities (Haimes, 2012; Harkins & Freed, 2017). The better prepared an organization is, the fewer exploitable vulnerabilities it will have.
Settlement Length	When an organization's information assets are compromised, the financial losses are significantly influenced by the time taken to restore operations or negotiate with attackers. Extended downtime disrupts services and diminishes revenue (Brown, et al., 2019). Moreover, prolonged recovery time would damage reputation, as customers may lose confidence in the organization's ability to safeguard their data or maintain reliable service (Anderson & Moore, 2006).
Connection	Organizations operate within complex networks of suppliers, partners, and customers, and a ransomware attack can disrupt this entire ecosystem (Robinson, et al., 2022; Cartwright & Cartwright, 2023). The ripple effects of external connections often extend the impact beyond the organization itself (Okafor, et al., 2022). For instance, a logistics company hit by ransomware might cause delays across a supply chain, leading to financial losses for all affected stakeholders. Cybersecurity is a process, not a product. Cyber risk management should permeate every part of business operations.
Total Cost	Direct losses, such as ransom payments, are immediate and quantifiable financial losses. Expenditure spending to regain access to their encrypted data, such as hiring experts to patch vulnerabilities, also add to the financial burden. Organizations that handle sensitive or regulated data may face fines for failing to secure the information safety (Robinson, et al., 2022; Nagar, 2024).

However, these factors are often examined in isolation, without integrated empirical modeling that quantifies their combined influence on financial loss outcomes.

To address these gaps, this study applies a GLM with a Gamma distribution and log link, supported by bootstrapping validation. By integrating socio-technical factors, the proposed model empirically quantifies ransomware loss severity using real-world data from the Advisen Cyber Loss Database (2018–2020). This approach bridges technical and managerial research streams, offering both academic insight and practical indicators for cyber risk management.

3. Quantitative Research

This study adopts an interdisciplinary framework to quantify ransomware losses by integrating socio-technical factors of cyber risk. Drawing from the Advisen Cyber Loss Database, we identify ransomware incidents from 2018 to 2020 and examine their financial outcomes. The selection of the Advisen dataset is based on its well-established reliability and its comprehensive incident-level financial records. Compared to other proprietary or survey-based sources, Advisen provides more comprehensive coverage across sectors and includes verified financial outcomes, which makes it particularly suitable for quantitative modeling of ransomware impacts.

We operationalize four categories of predictors: technology, preparedness, settlement length, and multi-entity connections. These variables are chosen based on prior literature and their theoretical relevance to organizational resilience.

The methodological goal is twofold. First, to determine how socio-technical factors shape ransomware loss severity. Second, to establish an empirical model that is generalizable and robust enough to inform cyber risk management practices.

3.1 Hypothesis Development

Based on literature review and industry evidence, we formulate four hypotheses regarding determinants of ransomware losses:

> *Hypothesis 1 (H1): Incidents involving servers generate higher loss severity compared to non-server incidents.*

Database server involvement is treated as a binary variable (1 = server affected, 0 = otherwise). Prior studies show that server-related incidents compromise critical infrastructure and data availability, amplifying costs.

> *Hypothesis 2 (H2): Lower preparedness is associated with higher ransomware loss severity.*

Preparedness is measured through insurance coverage ratios, representing the proportion of losses covered by cyber insurance. Organizations with low preparedness (uninsured or underinsured) are hypothesized to experience higher effective losses.

Hypothesis 3 (H3): Longer settlement duration is positively associated with greater ransomware loss severity.

Length refers to the duration (days) between incident onset and settlement. Prolonged negotiations, technical issues, or legal disputes can cause operational disruptions and reputational damage.

Hypothesis 4 (H4): Incidents involving multiple entities are associated with disproportionately higher losses.

This variable represents the number of affected entities in each incident (e.g., supply chain or partner firms). Interconnected organizations amplify cascading risks.

These hypotheses collectively reflect the socio-technical nature of ransomware attacks, acknowledging that financial impact arises not only from technical vulnerabilities but also from organizational and relational contexts.

3.2 Data Description

The dataset is derived from the Advisen Cyber Loss Database, a widely used comprehensive dataset. We filter cases from 2018–2020 that are explicitly classified as ransomware events. This temporal window is chosen for two reasons: (1) ransomware attacks increased substantially and evolved in sophistication during this period, and more complete reporting became available; (2) the version of the Advisen dataset used in this study does not include ransomware incidents beyond 2020. This limitation is further discussed in a later section. However, the primary contribution of this study is its methodological framework, which remains applicable and adaptable as more recent ransomware incident data become available.

- Dependent Variable:

Loss Severity (continuous): the reported financial impact of the ransomware incident. Because of right-skewed distribution, it is modeled using Gamma regression.

- Independent Variables:

Preparedness (ratio, 0–1): extent of cyber insurance coverage.

Server Binary (dummy, 0/1): whether servers were directly affected.

Settlement Length (days): duration of ransomware attacks.

Affected Entities (count): number of organizations impacted.

A summary of variable descriptions and coding is provided in Table 2.

Table 2: Description of Variables

Variable	Denoted	Type	Definition	Classification
Total Financial Loss	TOTAL	DV	The financial damage due to ransomware incident	Continuous
Database Server	SERVER	IV	Whether database servers were directly affected	Binary
Preparedness	PREPARED	IV	The level of support from cyber insurance	Continuous
Settlement Length	LENGTH	IV	The duration of ransomware attacks	Discrete
Affected Entities	AFFECTED	IV	The number of organizations impacted	Continuous

3.3 Model Specification

Ransomware loss data is highly skewed, with many moderate losses and a small number of extreme cases. To capture this distribution, we employ a GLM with Gamma regression and log link. This choice is appropriate

because: (1) the Gamma family is well-suited for continuous, positive-valued outcomes with right-skew (Ben-Assuli & Padman, 2020; Edwards, et al., 2016); (2) the log link ensures multiplicative interpretation, meaning predictors indicate proportional changes in expected losses (Ben-Assuli & Padman, 2020). Thus, Gamma regression with log link provides a statistically robust approach for modeling the ransomware loss. The model is expressed as:

$$\ln(E[\text{TOTAL_LOSS}_i]) = \beta_0 + \beta_1 \cdot \text{Preparedness}_i + \beta_2 \cdot \text{Server}_i + \beta_3 \cdot \text{Length}_i + \beta_4 \cdot \text{Affected}_i + \epsilon_i$$

To prepare the dataset, we retain outliers (extreme loss values beyond the 99th percentile) to preserve the distribution's integrity, while missing predictor values were imputed using median or mode as appropriate.

3.4 Validation Strategy

Given a limited sample size and skewed distributions, bootstrapping was used to evaluate the robustness of parameter estimates. Specifically, we conducted 100 resamples with replacement and recalculated coefficient estimates to generate empirical confidence intervals. This approach reduces reliance on asymptotic assumptions and ensures stability of findings under resampling variation.

All statistical analyses were conducted using R, a programming language designed for statistical analysis. Key packages include GLM for model fitting and boot for resampling. Data cleaning was performed using dplyr.

3.5 Summary

This methodology combines socio-technical insights with actuarial modeling to quantify ransomware loss severity. By focusing on four predictors—preparedness, server involvement, settlement length, and multi-entity connections—we establish an empirical framework that is both explanatory and predictive. The Gamma regression model with bootstrapping validation offers a statistically rigorous approach to understanding ransomware risks and provides practical indicators for insurers and policymakers.

4. Result Analysis

The regression results in Table 3 reveal several important insights regarding the factors influencing total reported losses from ransomware incidents.

Table 3: Gamma Regression Results (Log Link)

Variable	Estimate	Std. Error	z-value	p-value
Intercept	12.986	0.536	24.24	<0.001
PREPARED	11.002	11.133	0.99	0.323
AFFECTED	4.53e-11	7.09e-10	0.06	0.949
LENGTH	-0.0024	0.0006	-3.83	<0.001
SERVER	-53.018	0.536	-98.92	<0.001

Note: Dependent variable is TOTAL (total financial loss). Model uses a Gamma distribution with a log link. Significant predictors at $p < 0.05$ are in bold.

The intercept (estimate = 12.986, $p < 0.001$) represents the expected log-loss when all predictors are at their baseline levels, suggesting that even under minimal exposure conditions, ransomware incidents impose substantial financial burdens. The preparedness variable (PREPARED) shows a positive but statistically insignificant relationship with total loss, implying that the degree of insurance coverage does not consistently mitigate or amplify loss severity. This finding may reflect heterogeneity in policy structures, coverage limits, and reporting practices. Similarly, the coefficient for affected entities (AFFECTED) is near zero, indicating that the number of organizations linked to an incident does not meaningfully predict aggregate losses once other factors are controlled.

By contrast, the settlement length variable (LENGTH) has a significant negative relationship with total reported losses, which means prolonged negotiations may modestly reduce final loss amounts, possibly reflecting enhanced mitigation or recovery actions over time. The most notable effect emerges from database server involvement (SERVER). The large magnitude and statistical significance of this coefficient indicate that ransomware incidents involving database servers are associated with lower total reported losses, holding other factors constant. In other words, when ransomware attacks target database servers, organizations may experience comparatively reduced financial impacts relative to non-server incidents. This counterintuitive

finding may reflect underlying realities such as faster containment responses or earlier detection in server-related incidents. Regardless of the underlying causes, the result itself highlights the pivotal role of database server protection a mitigating potential financial losses from ransomware attacks.

4.1 Bootstrapping Validation

To assess the robustness of these results, Table 4 presents bootstrapping estimates derived from 100 resamples. The consistency of the coefficients across resampling reinforces the reliability of the model.

Table 4: Bootstrap Estimates for Gamma Regression Coefficients (100 Resamples)

Variable	Mean	Std. Dev.	2.5% CI	Median	97.5% CI
Intercept	12.412	1.005	10.656	12.893	13.858
PREPARED	14.093	20.514	-4.094	12.931	77.348
AFFECTED	3.08e-09	2.66e-08	-8.47e-09	1.47e-12	8.76e-09
LENGTH	0.0147	0.0306	-0.0066	-0.0019	0.0912
SERVER	-52.535	0.876	-53.858	-52.919	-50.967

The intercept remains stable (mean = 12.412, 95 % CI = 10.656–13.858), confirming a consistent baseline of ransomware loss magnitude. However, the preparedness variable (PREPARED) exhibits a large standard deviation (20.514) and an extremely wide confidence interval (–4.094 to 77.348), indicating instability and variability across resamples. This further supports the earlier conclusion that preparedness exerts no uniform or predictable influence on ransomware loss severity. Second, the affected entities variable (AFFECTED) retains an almost zero mean with a narrow confidence band centered around zero, reaffirming its negligible role in determining loss outcomes.

The bootstrap results confirm that the negative effects of both database server involvement (SERVER) and settlement length (LENGTH) are not only statistically significant in the original model but also consistently reproduced across resampled datasets. This consistency reinforces confidence in their robustness as key factors influencing total ransomware-related losses. LENGTH exhibits a generally negative trend, though with slightly greater variability, indicating a moderately consistent effect. SERVER shows strongly negative with a narrow confidence interval, reinforcing its role as a stable and influential predictor associated with lower reported losses.

In summary, the results from both the regression and the bootstrapping validation highlight that database server attacks are the dominant driver of financial loss in ransomware incidents, while settlement length has modest effect in the reduction of financial losses. Other socio-technical variables, such as preparedness and inter-organizational connections, contribute little explanatory power in this dataset. The findings suggest that organizations should prioritize strengthening database server security, data segregation, and backup protocols to mitigate the catastrophic impacts of ransomware. Additionally, organizations should recognize the value of coordinated incident response and post-incident recovery services and well develop incident response plans. Furthermore, the robust evidence from the bootstrapping analysis reinforces the reliability of these conclusions and demonstrates the effectiveness of the Gamma-regression framework in modeling skewed cyber-loss data.

5. Discussion and Future Work

The results provide important empirical insights into the mechanisms driving ransomware loss severity and demonstrate the value of quantitative cyber-risk modeling in technical security research. First, H1 emphasizes the influence of database server involvement on ransomware loss severity. The statistical significance of this variable highlights the central role of database servers in shaping financial outcomes and reinforces the need for prioritized protection of servers through segmentation, backup redundancy, and access-control hardening, as well as continuous monitoring for lateral movement within enterprise networks.

Regarding H3, the significant negative coefficient suggests that extended settlement period is associated with reduced financial losses. This pattern indicates that organizations benefiting from longer settlement periods may have more time to prepare negotiation strategies and implement effective mitigation measures to reduce potential losses. From a cybersecurity engineering standpoint, this highlights the importance of automated incident response systems, pre-approved recovery playbooks, and rapid decryption readiness testing to shorten response cycles without compromising long-term resilience.

H2 was not supported, as the preparedness variable was statistically insignificant in the regression model. Preparedness, operationalized through insurance coverage, did not significantly influence loss severity. This likely reflects the fact that insurance ratios do not capture underlying technical readiness, such as patching practices, endpoint protection, or offline backup quality. Similarly, H4 was not confirmed, as the number of affected entities showed no significant effect on losses. This result suggests that a simple count of affected organizations cannot capture the complexity of interorganizational dependencies, and may require more granular data, such as supply-chain topology, to capture the true impact.

These findings collectively demonstrate both the usefulness and limitations of the current dataset. While the Advisen Cyber Loss Dataset offers valuable and detailed information, its coverage is restricted to publicly reported or insured events, which may not include complex loss propagation patterns that occur within interconnected networks. For instance, the "affected entities" variable may not sufficiently capture the depth of supply-chain interdependencies or cross-organizational recovery coordination. Future work should incorporate more granular technical and organizational indicators (e.g. system topology, real-time telemetry, and incident response quality) to improve explanatory depth and extend the analysis to multi-source datasets to better support more comprehensive risk modeling.

In addition, the version in this study does not include ransomware incidents that occurred after 2020. The ransomware threat landscape has evolved considerably since the onset of the COVID-19 pandemic. Post-2020 developments, such as the emergence of more aggressive double-extortion and triple-extortion tactics, the rise of Ransomware-as-a-Service operations, and the integration of generative AI into cyberattacks, are therefore not captured in the current dataset. However, the analytical framework developed in this study remains applicable and adaptable to more recent datasets. As newer datasets become available, this methodology can be employed in future work to conduct updated analyses that reflect the latest trends in the ransomware landscape.

Methodologically, the study validates the applicability of Gamma regression with bootstrapping for cyber-loss analysis. This modeling approach effectively handles the heavy-tailed nature of ransomware data, reducing bias from extreme events while preserving interpretability. Bootstrapping further enhances robustness by confirming coefficient stability under repeated sampling, which is essential in cybersecurity datasets often characterized by limited incident reporting. Such techniques advance empirical rigor in cybersecurity research and bridge the gap between actuarial modeling and system security analytics.

To summarize, the findings not only quantify the economic impact of ransomware incidents but also provide a data-driven foundation for advancing technical cyber defense strategies. The integration of empirical loss modeling with operational cybersecurity practices can enhance predictive threat analysis, prioritize high-value targets for protection, and support the development of adaptive response mechanisms against evolving ransomware campaigns.

6. Conclusion

This study developed a data-driven framework to quantify ransomware loss severity using empirical incident data and a Gamma regression model validated through bootstrapping. By combining socio-technical variables with a statistical approach suited to positively skewed cyber-loss distributions, the analysis demonstrates that empirical modeling can enhance both academic understanding and practical defense of digital infrastructure.

The results confirm that ransomware incidents involving database server compromise are reliably associated with distinct financial impacts. This emphasizes the critical need for organizations to prioritize the protection of their information systems, particularly database servers, through robust access control, encryption, system segmentation, and timely patching of known vulnerabilities. In addition, the results reveal that shorter settlement durations are associated with reduced financial losses. This highlights the importance of well-designed incident response plans that facilitate rapid reaction and recovery. Moreover, continuous employee training, especially on social engineering tactics used by attackers, remains essential to reduce organizational exposure and enhance overall resilience.

Methodologically, this work validates the use of Gamma-based generalized linear modeling and bootstrapping as reliable tools for cybersecurity risk quantification. These techniques address the statistical challenges inherent in sparse and heavy-tailed cyber-incident datasets, offering a replicable path for empirical assessment of cyber-threat impacts that bridges data analytics and cyber defense operations.

From a technical defense perspective, the findings advocate for targeted hardening of server environments, implementation of redundant and isolated backups, and automation of recovery workflows to reduce both downtime and financial exposure. The framework can also support cybersecurity analytics by serving as a baseline for integrating AI-driven prediction or dynamic loss forecasting into security-operations platforms.

Future research should fuse this statistical foundation with machine learning based threat intelligence and real time telemetry analysis to construct adaptive, continuously learning risk models. Such integration will bridge quantitative cyber-loss modeling with active defense systems, advancing predictive resilience against the next generation of ransomware attacks.

Acknowledgements

This paper is based on parts of the author's doctoral dissertation, titled Safeguard Cyberspace in Ransomware Era: Risk Analysis & Cyber Insurance (University at Albany, 2025).

This work was made possible by the Advisen Cyber Loss Dataset. The authors thank Advisen Ltd. for providing this valuable resource.

Ethical Declaration: Ethical approval was not required.

AI Declaration: AI tools were not used in the creation of this paper.

References

AC Best, n.d. *Best's Market Segment Report: Ransomware and Aggregation Issues Call for New Approaches to Cyber Risk.* [Online] Available at: https://news.ambest.com/PR/PressContent.aspx?refnum=30762&altsrc=9[Accessed 6 June 2025].

Advisen Insurance Intelligence, 2015. *Cyber Risk Data Methodology for Insurance Risk Analysis.* [Online] Available at: https://www.advisenltd.com/wp-content/uploads/2015/04/cyber-risk-data-methodology-2015-04-30.pdf [Accessed 9 November 2024].

Anderson, R. et al., 2013. Measuring the Cost of Cybercrime. *The Economics of Information Security and Privacy,* pp. 265-300.

Anderson, R. & Moore, T., 2006. The Economics of Information Security. *Science,* 314(5799), pp. 610-613.

August, T., Dao, D. & Niculescu, M. F., 2019. *Economics of Ransomware.* s.l.:SSRN.

Ben-Assuli, O. & Padman, R., 2020. Trajectories of Repeated Readmissions of Chronic Disease Patients: Risk Stratification, Profiling, and Prediction.. *MIS Quarterly,* 44(1), pp. 201-227.

Brewer, R., 2016. Ransomware Attacks: Detection, Prevention and Cure. *Network Security,* 2016(9), pp. 5-9.

Brown, C. et al., 2019. Accounting for Business Adaptations in Economic Disruption Models. *Journal of Infrastructure Systems,* 25(1), p. 04019001.

Cartwright, A. & Cartwright, E., 2023. The Economics of Ransomware Attacks on Integrated Supply Chain Networks. *Digital Threats 4,* 56(14), pp. 1-56.

Dawson, M., Bacius, R., Gouveia, L. & Vassilakos, A., 2021. Understanding the Challenge of Cybersecurity in Critical Infrastructure Sectors. *Land Forces Academy Review,* 26(1), pp. 69-75.

Dekker, M. & Karsberg, C., n.d. *Technical Guideline on Incident Reporting,* s.l.: ENISA.

Dutta, K. & Perry, J., 2007. *A Tale of Tails: An Empirical Analysis of Loss Distribution Models for Estimating Operational Risk Capital,* s.l.: Federal Reserve Bank of Boston.

Edwards, B., Hofmeyr, S. & Forrest, S., 2016. Hype and Heavy Tails: A Closer Look at Data Breaches. *Journal of Cybersecurity,* 2(1), pp. 3-14.

Gordon, L. A. & Loeb, M. P. S. T., 2003. A Framework for Using Insurance for Cyber-risk Management. *Communications of the ACM,* 46(3), pp. 81-85.

Haimes, Y., 2012. Strategic Preparedness for Recovery from Catastrophic Risks to Communities and Infrastructure Systems of Systems. *Risk Analysis: An International Journal,* 32(11), pp. 1834-1845.

Harkins, M. & Freed, A. M., 2017. The Ransomware Assault on The Healthcare Sector. *JL & Cyber Warfare,* Volume 6, p. 148.

Hoferek, M. J. & Wilson, S. C., 2007. *Recovering from Database Recovery: Case Studies and The Lessons They Teach.* s.l., IEEE.

Kim, S. H. & Kwon, J., 2019. How do EHRs and a Meaningful Use Initiative Affect Breaches of Patient Information?. *Information Systems Research,* 30(4), pp. 1184-1202.

Li, W., Leung, A. & Yue, W., 2023. Where is IT in Information Security? The Interrelationship among IT Investment, Security Awareness, and Data Breaches. *MIS Quarterly,* 47(1), pp. 317-342.

Martin, A. & Collier, J., 2020. Beyond awareness: Reflections on Meeting the Interdisciplinary Cyber Skills Demand. *Cyber Security Education,* pp. 55-73.

McIntosh, T. et al., 2021. Ransomware Mitigation in The Modern Era: A Comprehensive Review, Research Challenges, and Future Directions.. *ACM Computing Surveys (CSUR),* 54(9), pp. 1-36.

Mukhopadhyay, A. et al., 2019. Cyber Risk Assessment and Mitigation (CRAM) Framework Using Logit and Probit Models for Cyber Insurance. *Information Systems Frontiers,* 21(5), pp. 997-1018.

Mukhopadhyay, A. & Jain, S., 2024. A Framework for Cyber-risk Insurance Against Ransomware: A Mixed-method Approach. *International Journal of Information Management,* Volume 74, p. 102724.

Nagar, G., 2024. The Evolution of Ransomware: Tactics, Techniques, and Mitigation Strategies. *Valley International Journal Digital Library,* pp. 1282-1298.

Neoaz, N., 2024. Cybersecurity and Information Assurance: Bridging the Gap. *International Journal of Social, Humanities and Life Sciences,* 2(1), pp. 37-46.

Okafor, C., Schorlemmer, T., Torres-Arias, S. & Davis, J., 2022. *SoK: Analysis of Software Supply Chain Security by Establishing Secure Design Properties.* s.l., ACM.

OZ, H. O., Aris, A., Levi, A. & Uluagac, A., 2022. A Survey on Ransomware: Evolution, Taxonomy, and Defense Solutions. *ACM Computing Surveys (CSUR),* 54(11), pp. 1-37.

Rassam, M., Maarof, M. & Zainal, A., 2017. Big Data Analytics Adoption for Cybersecurity: A Review of Current Solutions, Requirements, Challenges and Trends. *Journal of Information Assurance & Security,* 12(4).

Robinson, A., Corcoran, C. & Waldo, J., 2022. *New Risks in Ransomware: Supply Chain Attacks and Cryptocurrency,* s.l.: s.n.

Samonas, S., Dhillon, G. & Almusharraf, A., 2020. Stakeholder Perceptions of Information Security Policy: Analyzing Personal Constructs. *International Journal of Information Management,* Volume 50, pp. 144-154.

Tsen, E., Ko, R. & Slapnicar, S., 2022. An Exploratory Study of Organizational Cyber Resilience, Its Precursors and Outcomes. *Journal of Organizational Computing and Electronic Commerce,* 32(2), pp. 153-174.

First Steps, Lasting Impact: Platform-Aware Forensics for the Next Generation of Analysts

Vinayak Jain[1], Sneha Sudhakaran[1] and Saranyan Senthivel[2]
[1]Florida Institute of Technology, Melbourne, USA
[2]DXC Technology

vjain2023@my.fit.edu
ssudhakaran@fit.edu
ssenthivel2@dxc.com

Abstract: The reliability of cyber forensic evidence acquisition is strongly influenced by the underlying operating systems, Windows, macOS, and Linux - due to inherent variations in file system structures, encryption protocols, and forensic tool compatibility. Disk forensics, one of the most widely used techniques in digital investigations, faces distinct obstacles on each platform. Windows, with its predominantly NTFS and FAT file systems, typically supports reliable disk imaging and analysis through established tools such as FTK Imager and Autopsy/Sleuth Kit. However, encryption features frequently pose challenges to evidence acquisition. Conversely, Linux environments, which rely on file systems like ext4 and XFS, generally offer greater transparency, yet the transient nature of log retention often complicates forensic analysis. In instances where anti-forensic strategies—such as encryption and compression— render traditional disk forensics insufficient, memory forensics becomes crucial. While memory forensic methodologies demonstrate robustness across Windows and Linux platforms forms through frameworks like Volatility, platform-specific difficulties persist. Memory analysis on Linux systems benefits from tools like LiME, snapshot utilities, and dd for memory acquisition; nevertheless, live memory acquisition on Linux can still present challenges. This research systematically assesses both disk and memory forensic acquisition techniques across samples representing Windows and Linux systems. By identifying effective combinations of forensic tools and configurations tailored to each operating system, the study aims to improve the accuracy and reliability of evidence collection. It further evaluates current forensic tools and highlights a persistent gap: consistently assuring forensic input reliability and footprint integrity. By integrating static analysis with temporal performance evaluations, this study emphasizes the importance of developing standardized forensic procedures to ensure the integrity and admissibility of evidence across diverse platforms. Establishing such standardized methodologies is imperative to maintaining the credibility and reliability of cyber forensic investigations in heterogeneous operating system environments.

Keywords: Digital forensics, Evidence acquisition, Memory forensics, Disk forensics, Anti-Forensic techniques, Forensic tool reliability

1. Introduction

The reliability of cyber forensic evidence acquisition is strongly influenced by the operating system under investigation, specifically Windows, macOS, or Linux, due to fundamental differences in file system structures, encryption mechanisms, and the compatibility of forensic tools. Among the most widely adopted approaches, disk forensics provides a foundation for digital investigations; however, it encounters distinct challenges across various platforms. Windows systems, which predominantly employ NTFS and FAT file systems, are well supported by established tools such as FTK Imager (FTK Imager, 2000) and Autopsy/Sleuthkit(Autopsy,2001; Sleuthkit,2001) . Still, built-in encryption features often complicate the acquisition of reliable information. Linux environments, in contrast, commonly use ext4 and XFS file systems that offer greater transparency. However, the ephemeral nature of system logs and varying retention policies creates obstacles to comprehensive evidence collection. When traditional disk-based approaches are hindered by encryption, compression, or other anti-forensic techniques, memory forensics becomes indispensable. Memory acquisition and analysis frameworks, such as Volatility, offer cross-platform capabilities; however, platform-specific hurdles persist. On Linux, for example, tools such as LiME (LIME, 2012)] and raw memory acquisition with dd support forensic collection, though live memory acquisition continues to pose technical difficulties(Forensics College, 2023).

This research systematically examines both disk and memory forensic acquisition techniques across representative Windows and Linux systems. By identifying the most effective combinations of tools and configurations tailored to each platform, the study seeks to enhance the accuracy and reliability of digital evidence collection. Beyond evaluating practical performance of forensic tools, it highlights a persistent research gap: ensuring the consistency, reliability, and integrity of forensic inputs across heterogeneous environments. To address this challenge, the study integrates static analysis with temporal performance assessments, underscoring the need for standardized forensic procedures. Such standardization is critical to safeguarding the credibility and admissibility of cyber forensic evidence in contemporary investigations.

Detecting anti-forensic techniques ensures that attempts at obfuscation or deletion do not obscure critical evidence. Building on the preceding analysis, we aim to identify reliable and effective forensic methods by conducting an empirical evaluation across a range of simulated case studies. Our contribution in this paper marks two primary research objectives:

Research Objective 1: To systematically evaluate the effectiveness of different forensic techniques across simulated case studies, identifying those that yield the most reliable and trustworthy results.

Research Objective 2:To develop a set of best practices for assessing the credibility and accuracy of forensic methods through empirical analysis, ensuring consistency in real world applications.

In this paper, Section 2 explains the related work, Section 3 describes the design methodology, Section 4 presents results, and Section 5 concludes with future work.

2. Literature Survey

2.1 Cross-Platform Forensic Analysis for Cybercrime Investigations

Digital, computer, and cyber forensics have emerged as critical disciplines in modern criminal investigations, driven by the widespread proliferation of computing devices. This growth has introduced both opportunities and challenges for investigators. As cybercrime becomes increasingly sophisticated, forensic methodologies must evolve to address the diverse array of operating systems encountered in investigations. This literature survey focuses on empirical approaches to forensic analysis across desktop operating systems – Windows, Linux, and macOS, for future work reference, with a specific emphasis on the challenges posed by malware infections. The importance of cross-platform forensic proficiency has increased significantly as

Cyber criminals increasingly employ multiple operating systems in their operations. Recent findings suggest that approximately 43% of sophisticated cyberattacks involve multiple operating systems, either as targets or as platforms for attack (EBSCO, 2023). This multi-platform strategy presents distinct challenges for forensic investigators, who must possess expertise across various operating environments, each with unique file systems, memory management practices, and evidence artifact locations. This survey synthesizes current research, methodological frameworks, and empirical findings to provide a comprehensive understanding of reliable forensic analysis techniques applicable across diverse operating systems. Particular emphasis is placed on memory dumps and disk images as primary evidence sources, with close examination of how the presence of malware impacts the acquisition and analysis of these forensic artifacts. Commercial forensic tools are widely used for evidence acquisition and analysis, with varying cross-platform capabilities. For instance, EnCase and X-Ways Forensics are regarded for windows(registry artifacts, NTFS), while BlackLight targets APFS on macOS and Magnet AXIOM provides cross-platform timeline analysis (ForensicColleges, 2023;Magnet,2024). One common drawback of commercial forensic tools is their delay in supporting new operating system versions. Studies have shown that it often takes several months for these tools to fully adapt to major updates, with macOS updates typically experiencing the longest delays compared to Windows and Linux distributions. This delay can impede timely investigations, particularly when dealing with cutting-edge technologies or recently released operating systems (Forensic Focus, 2023).

2.1.1 Open-Source forensic tools

Open-source forensic tools are becoming increasingly popular as cost-effective and flexible alternatives to commercial solutions. Tools such as The Sleuth Kit offer strong support for Linux file systems like EXT4, though they often lack comprehensive compatibility with macOS's APFS(Sleuthkit, 2001). In cloud environments, Google Cloud Forensics Utils integrates open-source frameworks such as Plaso and Log2Timeline to streamline log analysis in hybrid environments(CloudForensicutil, 2023). For memory forensics, Volatility has established itself as a leading framework, providing robust capabilities for examining memory artifacts on both Windows and Linux platforms (Ligh,2014;Volatility,2014). Open-source solutions frequently demonstrate faster adaptation to emerging technologies than their commercial counterparts, driven by active community contributions.

This adaptability makes them particularly valuable for investigating unusual configurations, addressing novel file systems, and responding to newly developed platforms (BlueVoyant, 2025 ; ForensicFocus, 2023). The above literature has helped us to inform tool selection for acquisition and analysis across platforms. Digital forensic investigations are based on established frameworks to ensure systematic and methodical handling of evidence. Among the most prominent is the DFRWS Investigative Model, which comprises six well-defined phases: identification, preservation, collection, examination, analysis, and presentation (Palmer, 2001). This framework

emphasizes scientific rigor and legal admissibility, serving as a cornerstone for modern investigative methodologies (Infosec, 2016). Furthermore, the ACPO Good Practice Guide builds on these phases, placing particular emphasis on the chain-of-custody protocols and the validation of forensic tools (Colvee,2023). There are many other tools that are specific to mobile device forensics like AmpleDroid, DroidScraper(Ali-Gombe et al, 2017; Ali-Gomber et al,2023; Sudhakaran,2022; Sudhakaran et al,2022; Sudhakaran et al, 2020). Though there are multiple open source tools developed in forensics the creation of such tools are also to be made aware to public as we see that there are surveys conducted which reveal that usage of such tools for people from academia is not easy and is not much accessible as shown in the work presented in(Sudhakaran et al, 2024).

2.2 Acquisition of Evidential Data in a Trusted Environment

Once we identify that the forensically acquired dump has a role in anti-forensic data,a forensic investigator must acquire runtime or volatile data. This enables the effective acquisition and analysis of volatile data for efficient evidence recovery. The acquisition of volatile data is one of the most critical and technically challenging steps in digital forensics. Preserving ephemeral evidence before it is lost is paramount, and memory forensics plays a central role in this effort. By analyzing system RAM, investigators can recover valuable information that is not present in persistent media. Tools such as Belkasoft Live RAM Capturer are widely used in Windows systems due to their ability to operate at the kernel level and bypass anti-debugging or anti-dumping measures ensuring reliable memory extraction (ForensicsFocus,2023; Belkasoft,2023).

For Linux, LiME (Linux Memory Extractor) provides full memory captures while limiting interactions between user and kernel space processes, thus maintaining forensic soundness. However, LiME can face challenges when compiling kernel modules in various Linux distributions [48, 49]. On macOS, tools such as OSXPMem and MacQuisition are commonly used for capturing volatile memory. Although effective, these tools

can encounter difficulties with Apple's advanced memory protection features, sometimes resulting in system crashes that risk data loss if not properly managed (Naval,2017;Ponder,2017). Delays in acquisition increase the chance of losing critical artifacts, as even minimal tool activity can modify memory contents. Despite these complications, volatile memory analysis remains indispensable, often uncovering encryption keys, login credentials, and other vital evidence Forensicswiki, 2023;Belkasoft,2023; Ponder,2017). Hash verification is a cornerstone of verifying evidence integrity. Although MD5 and SHA-1 are still widely used, studies have highlighted their susceptibility to collision attacks, prompting a shift towards stronger hashing algorithms, such as SHA-256 (TechFusion,2024). Adopting these more robust algorithms is recommended for verifying all digital evidence, regardless of the operating system's source. Specialized forensic distributions, including SANS SIFT, Kali Linux, and CAINE, have gained prominence for creating standardized analysis environments. Each distribution offers unique strengths. SIFT is known for its broad compatibility with Windows, Linux, and macOS, while CAINE excels at integrating legal documentation workflows, but provides limited support for macOS artifacts. Cloud-based platforms are emerging as essential tools in digital investigations. Solutions like Cado Security automate evidence collection in cloud environments such as AWS and Azure, significantly reducing analysis time compared to traditional methods (APUS,2024; Sentinel,2024). Timeline reconstruction tools, such as Magnet AXIOM, correlate events across platforms to establish comprehensive attack chronologies. Challenges such as clock synchronization discrepancies between systems (e.g., Windows vs. Linux) complicate cross-platform timelines but can be mitigated through rigorous NTP implementation. Visualization techniques such as temporal heat maps improve the recognition of attack patterns, reducing analysis time, and enhancing investigative efficiency (SalvationData,2023). Correlating digital artifacts across platforms is essential for modern investigations. Tools such as Cado Security map cloud activity to on-premises evidence, linking AWS/Azure logs with local user actions. Graph-based analysis models files, processes, and users as nodes with interactions as edges, revealing hidden connections in multi-platform attacks. Behavior-based correlation focuses on attacker objectives rather than platform-specific methods, improving the detection of advanced threats tailored to specific environments (CadoSecurity,2023).

Our work responds to a critical gap in the literature, namely the limited availability of beginner-friendly, systematic instructions for configuring and understanding formal digital forensic analysis environments.

3. Design Methodology

This part presents the empirical framework adopted to evaluate the reliability of forensic methods across Windows and Linux platforms. The methodology as shown in Fig 1 is structured into three modules: Case Study Creation, Reliable Evidence Acquisition, and Reliable Analysis for Investigation.

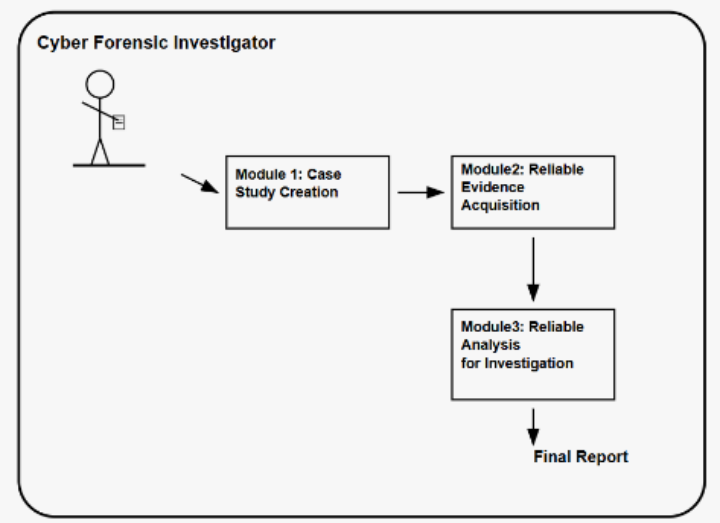


Figure 1: Environment Setup

3.1 Module 1 – Case Study Creation

All experiments were performed in standardized virtual environments (VMware and VirtualBox) on an Intel i5-7200U host system to minimize external artifacts. Four operating systems—Windows 11, Windows 10, Ubuntu 24, and Ubuntu 20—were configured with 4GB and 8GB RAM and 80GB storage and is depicted in Fig 2. Simulated user activity (browsing, file handling, and media use) created realistic conditions. Baseline clean-state memory and disk images were collected for comparison.

Operating System	Version	RAM Configuration	Storage Capacity
Windows	Windows 11	4GB / 8GB	80GB
Windows	Windows 10	4GB / 8GB	80GB
Linux	Ubuntu 24	4GB / 8GB	80GB
Linux	Ubuntu 20	4GB / 8GB	80GB

Figure 2: Operating Systems and Their Configurations

A total of 32 case studies per platform were designed to evaluate forensic reliability under increasing infection complexity. Malware included Trojans (CrossRAT, veryfun.exe, icon_dance.exe; dakkatoni, gosh, brootkit) and ransomware (cerber, cryptolocker, WannaCry; ransomexx, monti, gonnacry) from open Internet randomly selected. Infection states ranged from clean to single, double, and triple infections, allowing for the analysis of evidence persistence and malware evolution.

3.2 Module 2 – Reliable Evidence Acquisition

Stage C0: Each malware type was installed on a clean VM and executed for at least 10minutes (longer for complex samples) before acquisition. **Stage C1:** One malware instance (same category) was executed, then acquired. **Stage C2:** Two instances (same category) were executed, then acquired. **Stage C3:** Three instances were executed in isolated environments, then acquired to prevent spread while maintaining realism. During these stages, systems remained operational with regular user-like activities, including web browsing, document file access, and application launches and closures. This approach allowed for the assessment of the

persistence of evidence and the evolution of malware behavior over time in different operating systems. Memory dumps were collected with FTK Imager (Windows) and LiME (Linux). Disk images were acquired with FTK Imager (Windows) and dd (Linux). Each acquisition was repeated under both RAM configurations, and cryptographic hashes were used to ensure the integrity of the evidence.

Sections 3.1 and 3.2 satisfy Research Objective 1.

3.3 Module 3 – Reliable Analysis for Investigation

Table 1: Benefits of Open Source vs Closed Source Memory Forensic tools

Aspect	Open-Source Tools (Volatility, Rekall, LiME)	Commercial Tools (Magnet Axiom, Belkasoft, X-Ways)
Ease of Use	Harder to learn; requires commands and technical skills and no cost benefit	Easy to use; GUI-based and beginner-friendly and expensive
Human Effort Needed	Higher effort was needed for analyst to interpret results manually	Low effort from human and tool guides the user and highlights important findings
Flexibility	Very flexible - customizable and automation scripting was possible	Less flexible and fixed workflows
Transparency	Very clear to show raw data and plugin output	Less transparent and internal processes hidden
Investigation Speed	Slow for beginners fast for experts	Fast for most users
Best For	Research, teaching, advanced deep-dive analysis	Law enforcement, quick triage, non-experts

This experimental study provides a systematic set of best practices for evaluating forensic capabilities in diverse computing environments. By implementing various test scenarios across different operating systems with varying hardware configurations, the research enables empirical assessment of the effectiveness of forensic tools in real-world scenarios. We created analysis scenarios for trojans and ransomware and evaluated their effectiveness under time- and resource-dependent conditions across various tools. We conducted experiments with multiple time and resource-dependent analyses across different tools and ensured the scenarios were achieved. The Section 3.3 satisfies Research Objective 2. Overall the ideal case for users to start learning forensic analysis would be beneficial for long run with Open source tools as shown in Table 1

3.3.1 Malware-specific analysis challenge

Trojans may hide processes, inject into legitimate processes, and establish persistent connections detectable in memory forensics but harder to observe in disk forensics. Ransomware will likely create readily identifiable artifacts in both disk and memory forensics due to extensive file system modifications and encryption operations.

3.3.2 Operating system-specific challenges may include

Windows systems exhibited complications with registry analysis and memory address translation. Linux systems presented challenges with different kernel versions affecting memory structure mapping.

4. Evaluation and Effectiveness Assessment

The empirical analysis highlights the impact of trojans and ransomware on Windows and Linux systems. We conducted memory analysis with Volatility on dumps acquired from both platforms.

4.1 Windows Case Study Findings: Impact of Trojans and Ransomware on Forensic Analysis

4.1.1 Memory forensic analysis on trojan and ransomware case investigations

Volatility plugin	C0 Clean state		C1 ransomware infection		C2 ransomware infection		C3 ransomware infection	
	8gb	4gb	8gb	4gb	8gb	4gb	8gb	4gb
Windows.pslist	131	133	132	141	125	136	120	122
Windows.psscan	132	139	131	143	135	141	123	144
Windows.pstree	130	133	132	140	133	138	120	135
Windows.malfind	22	24	14	11	19	11	5	7

Figure 3: Windows 10 ransomware malware process count descriptions

Based on these findings in Fig 3 and 4, we recommend prioritizing 8 GB Windows 10/11 environments for memory forensics in Trojan-infected systems to maximize evidence retention and analysis efficiency, while acknowledging the potential but limited utility of 4GB systems where hardware constraints apply. Greater memory capacity improves artifact retention and accommodates system-level constraints such as page swapping.

4.1.2 Disk forensic analysis on trojan and ransomware case investigation

Analysis results	C0 clean state	C1 Trojan infection	C2 Trojan infection	C3 Trojan infection	C1 Ransomware	C2Ransomware	C3Ransomware
Encryption Suspected	4	3	2	2	11	9	8
Extension Mismatch Detected	254	261	260	265	255	383	215
Suspicious items	258	264	263	268	267	393	223
Bad items	0	0	0	0	0	0	0
Interested Item (zip bomb)	0	0	1	1	1	1	0
Recycle bin	19	19	19	19	19	19	19

Figure 4: Analysis Results of process count descriptions for Various Infections of Windows 10 (8 GB)

Case studies involving Trojan and ransomware infections revealed significant variations across Windows versions. We conducted the analysis using Autopsy 4.21.0, focusing on suspected encryption activities and extension mismatches, as well as identifying suspicious items, bad items, potential zip bombs, and artifacts within the Recycle Bin. Based on empirical results, we conclude that Windows 11 demonstrates a stronger capability to resist and detect malware compared to Windows 10 within the context of our disk forensics analysis. We observed a slight increase in detected artifacts from the C1 to C3 malware infection states, except for the C3 ransomware infection stage.

4.2 Linux Case Study Findings: Impact of Trojans and Ransomware on Forensic Analysis

4.2.1 Memory forensic findings across linux systems

Examining memory dumps from various Linux environments infected with trojan malware revealed notable differences in forensic visibility as shown in Figures 5 and 6. We conducted our analysis using the Volatility 3 framework with Linux-specific plugins, including linux.pslist, linux.pstree, and linux.pidhashtable, to identify distinct malware behaviors. Additionally, we utilized the linux.elf plugin to detect suspicious or malicious ELF files during memory forensics.

Volatility plugin	C 0 Clean state		C1 trojan infection		C2 trojan infection		C3 trojan infection	
	8gb	4gb	8gb	4gb	8gb	4gb	8gb	4gb
linux.pslist	148	153	151	137	136	129	145	137
linux.pidhashtable	151	155	140	143	143	144	152	136
linux.pstree	147	153	149	91	136	121	150	139

Figure 5: Ubuntu 24 trojan malware process count descriptions

Figure 5 above presents the number of processes identified by each plugin during the analysis of Ubuntu 24.04 and Ubuntu 20.04 systems with 8 GB and 4 GB RAM configurations infected with trojan malware. Our analysis revealed that trojan malware infections exhibited similar behavioural patterns in both Linux operating systems examined. However, we observed a notable difference in the time required to collect memory dumps between the systems. Specifically, Ubuntu 24.04 required, on average, 5–8% longer to complete memory dump acquisition compared to Ubuntu 20.04.

Volatility plugin	c 0 Clean state		C1 ransomware infection		C2 ransomware infection		C3 transomware infection	
	8gb	4gb	8gb	4gb	8gb	4gb	8gb	4gb
linux.pslist	144	134	154	112	x	x	x	x
linux.pidhashtable	150	138	151	134	x	x	x	x
linux.pstree	148	138	154	115	x	x	x	x

Figure 6: Ubuntu 24 Ransomware malware process count descriptions

Figures 5 and 6 clearly shows the ransomware infections created substantially different forensic patterns compared to trojans. We have used the same procedure to take the memory dumps as in the trojan infection case and the same volatility plugins for memory forensic analysis. Our observations indicate that ransomware variants such as monti and gonnacry exhibit a significantly higher potential to harm systems compared to trojan malware. Ransomware infections result in a greater degree of system corruption, actively modifying system files in ways that can interfere with the memory acquisition process, whereas trojan infections generally do not cause such extensive file system alterations.

4.2.2 Disk forensic findings across linux systems

Based on our analysis across both system configurations (8GB and 4GB RAM) running the same Linux operating system, we found that variations in RAM capacity did not produce a substantial difference in the results of disk

forensics. Compared to trojan malware, ransomware infections in Linux systems exhibited an approximate 8–15% increase in suspicious activity. trojan infections, by contrast, showed a 3–5% increase in activity as the number of trojan samples introduced into the system increased. Analysis of the Recycle Bin revealed that, even after ransomware encryption, disk forensics remained effective in recovering deleted data artifacts from the Recycle Bin.

4.3 Errors and Complications in Forensic Analysis

During our comprehensive analysis across operating systems, we encountered various errors and complications that impacted forensic processes and outcomes. These challenges arose across both Windows and Linux environments and were influenced by factors related to the host machines used for case studies as well as the virtual machine platforms employed, including VMware Workstation and Oracle VirtualBox.

4.3.1 Operating system complications

Many types of malware fail to execute due to compatibility issues between the operating system and the malware's executable format. For example, Linux systems support ELF (Executable and Linkable Format) files, while Windows systems support .exe files. This disparity makes it challenging to generalize how the same malware type affects different operating systems.

On Windows 11, after acquiring forensic data on an 8 GB RAM configuration for trojan and ransomware samples, the virtual machine displayed errors such as triggering the automatic repair process. Even after several hours, the Windows 11 VM host failed to restart. Attempts to resolve the issue using a new ISO image

and switching to Oracle VirtualBox were unsuccessful, with the Windows 11 VM failing to start in both VMware Workstation and VirtualBox. Consequently, we could not acquire forensic data for the 8 GB configuration on Windows 11, an issue that did not occur on Windows 10 VMs. The host system occasionally experienced sudden crashes during forensic acquisition and analysis, manifesting as a blank white screen or system lockup/freezes.

Extended periods without restarting the host machine led to delays when opening virtual machines on both VMware Workstation and VirtualBox. Ubuntu 24.04, in particular, exhibited signs of system overload, failing to open immediately as Ubuntu 20.04 did. In these cases, restarting the entire host machine was necessary to resume work on Ubuntu 24.04, whereas for Ubuntu 20.04, closing and reopening the VM itself was sufficient.

4.3.2 Memory forensic complications

Volatility is available as Versions 2 and 3. Because Volatility 2 does not support Windows 11, we used Volatility 3 for all Windows 10/11 memory analyses.

When analyzing memory dumps from Windows 11, Volatility may display symbol table and profile errors. If using an older version, such as Volatility 2.7.0, upgrading to the latest version is recommended to resolve these errors. It is essential to ensure that all required Python libraries are installed in the environment otherwise, ModuleNotFoundError may occur. Additionally, specifying the correct path to the memory dump is critical, as incorrect paths will result in FileNotFoundError.

Malware infections, particularly ransomware, can corrupt the system's OS symbol table. In our Linux OS case studies, we observed that at the C2 and C3 infection stages, forensic analysis often failed due to malware-induced encryption and kernel corruption, resulting in corrupted memory dumps. Consequently, Volatility was unable to process these memory dumps for analysis.

4.3.3 Disk forensic complications

In disk forensics, autopsy requires an average of 2–3 hours to process a given disk dump. During this process, it is recommended that the host machine remain idle, as concurrent activities may cause the application to freeze or enter a processing loop, necessitating a complete restart of the analysis to obtain accurate forensic data.For Windows systems, the use of FTK Imager during disk dump acquisition is recommended. This tool can export a structured list of files and directories from an acquired image to Excel-compatible formats (CSV or .xlsx), facilitating forensic reporting, timeline reconstruction, and the identification of suspicious files, such as those with unusual extensions, large sizes, or hidden attributes.

For Linux systems, selecting the correct file-system path (e.g., /dev/sda) is essential during acquisition. Failure to specify the correct disk path will prevent Autopsy from accurately reading and processing the disk dump. Severe disk corruption or ransomware overwrites can hinder Autopsy's ability to carve and recover files from disk images successfully. Malware often hides in directories such as /tmp/.xyz or /var/.abc. These hidden files

may be overlooked unless explicitly searched during analysis. Files with extensions such as .jpg or .txt may, in reality, be executable or encrypted files. Malware infections may also generate duplicate files containing random data, which share identical hash values, thereby complicating file integrity analysis.

4.4 Limitations and Future Research Directions

While this research provides comprehensive information on the behavior of malware in Windows and Linux operating systems, several limitations must be acknowledged. The use of virtual machines, while essential for consistency and isolation, may not fully replicate physical hardware behavior(e.g., memory management, storage performance). Future studies should consider testing on physical systems to validate these findings. Furthermore, while the selected malware samples represent common cross-platform threats, they do not capture the full diversity of malware types, particularly those designed for specific operating systems. Future research should incorporate a broader range of targeted malware samples to enhance generalizability. Reliance on Volatility 3 and Autopsy 4.21.0 introduces potential tool-specific biases. Comparing results across multiple forensic tools would help distinguish genuine platform behaviors from tool-induced artifacts, thus enhancing the robustness and reliability of forensic analysis outcomes.

5. Conclusion

This research presents a systematic empirical analysis of reliability challenges in cyber forensic investigations across Windows and Linux environments, using a controlled experimental framework with virtual machines to ensure hardware consistency. Four operating systems were evaluated under varied malware infection scenarios, system configurations, and acquisition timelines, revealing that operating system architecture, memory resources, malware type, and the timing of forensic acquisition significantly affect evidence visibility and degradation. By incorporating both memory and disk forensics with standardized procedures and platform-specific tools, the study provided consistent and credible cross-platform evaluations. Key contributions include a cross-platform forensic comparison under uniform malware conditions, quantification supporting a practical minimum of 8 GB RAM for reliable analysis; identification of distinct forensic patterns in trojan and ransomware infections, empirically grounded timelines for evidence degradation, and documentation of tool-specific error behaviors across platforms to guide practitioners. The results further suggest that encrypted content demonstrates resilience against ransomware attacks, highlighting encryption as a defensive strategy. Nevertheless, limitations arise from using virtual machines, which may not fully replicate physical hardware behaviors, the restricted set of malware samples, and potential tool-specific biases from Volatility 3 and Autopsy 4.21.0. Future work will expand this analysis to macOS environments and incorporate a broader range of targeted malware to strengthen generalizability.

AI declaration: Artificial intelligence tools (OpenAI ChatGPT, GPT-5) were employed solely to improve grammar, readability, and formatting of the manuscript text. The authors take full responsibility for the scientific content, data interpretation, and conclusions. The authors also used this tool response as this work mainly focus on the usage of Gen AI tool like Chat Gpt with Human forensic investigators for better efficacy over cyber forensic investigation. All outputs from AI tools were carefully reviewed and edited by the authors to ensure accuracy and originality.

Ethics Declaration: This study was conducted in accordance with established ethical and professional standards. No human participants, personal data, or sensitive information were involved; all datasets and experiments were performed on controlled test environments or publicly available resources. The work is intended solely for research and educational purposes in cybersecurity and does not promote or support malicious use.

References

504ensics Labs (2012) *LiME – Linux memory extractor*. Available at: https://github.com/504ensicsLabs/LiME

Ali-Gombe, A., Sudhakaran, S., Case, A. and Richard III, G.G. (2019) 'DroidScraper: A tool for Android in-memory object recovery and reconstruction', *Proceedings of the 22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019)*, pp. 547–559.

Ali-Gombe, A., Sudhakaran, S., Vijayakanthan, R. and Richard III, G.G. (2023) 'cRGB_Mem: At the intersection of memory forensics and machine learning', *Forensic Science International: Digital Investigation*.

APUS (2024) *What is digital forensics?* Available at: https://www.apu.apus.edu/area-of-study/information-technology/resources/what-is-digital-forensics/

Autopsy Digital Forensics/Sleuth Kit Labs (n.d.) *Autopsy*. Available at: https://www.autopsy.com

Belkasoft (2002) *Belkasoft RAM Capturer*. Available at: https://belkasoft.com/ram-capturer

BlueVoyant (2025a) *Get started with these 9 open-source tools*. Available at: https://www.bluevoyant.com/knowledge-center/get-started-with-these-9-open-source-tools

BlueVoyant (2025b) *Understanding digital forensics: Process, techniques and tools*. Available at: https://www.bluevoyant.com/knowledge-center/understanding-digital-forensics-process-techniques-and-tools
Cado Security (2023) *Top free and open-source forensics tools*. Available at: https://www.cadosecurity.com/blog/top-free-and-open-source-forensics-tools
Cloud Forensics Utils (2023) *Google cloud-forensics-utils*. Available at: https://github.com/google/cloud-forensics-utils
ColVEE (2023) *Computer forensic investigation process – first response*. Available at: https://tell.colvee.org/pluginfile.php/3104/mod_resource/content/4/Unit204.pdf
EBSCO (2023) *Cyber crimes and forensics: Research starters*. Available at: https://www.ebsco.com/research-starters/science/cyber-crimes-and-forensics
Forensic Focus Forums (2023) *How important is cross-platform support for forensic tools?* Available at: https://www.forensicfocus.com/forums/general/how-important-is-cross-platform-support-for-forensic-tools/
Forensics Colleges (2023) *Guide to digital forensics tools*. Available at: https://www.forensicscolleges.com/blog/resources/guide-digital-forensics-tools
Forensics Wiki (2023) *Belkasoft live RAM capturer*. Available at: https://forensics.wiki/belkasoft_live_ram_capturer/
FTK Imager (2000) *Forensic Toolkit (FTK) Imager*. AccessData/Exterro. Available at: https://accessdata.com/product-download/ftk-imager-version-4-2-1
Infosec Institute (2016) *Digital forensics models*. Available at: https://www.infosecinstitute.com/resources/digital-forensics/digital-forensics-models/
Jeffcoat, C. (2023) *Importance of volatile data capture in digital forensics*. Available at: https://www.linkedin.com/pulse/importance-volatile-data-capture-digital-forensics-jeffcoat-cissp
Ligh, M.H., Case, A., Levy, J. and Walters, A. (2014) *The art of memory forensics: Detecting malware and threats in Windows, Linux, and Mac memory*. Wiley.
Magnet Forensics (2024) *Digital forensics tools: The ultimate guide 2024*. Available at: https://www.magnetforensics.com/blog/digital-forensics-tools-the-ultimate-guide-2024/
Naval Postgraduate School (2017) *Mac memory acquisition and analysis using OSXPMem and Volatility*. Available at: https://faculty.nps.edu/ncrowe/oldstudents/icdf2c17_leopard.htm
Palmer, G. (2001) 'A road map for digital forensic research', *First Digital Forensic Research Workshop (DFRWS)*.
Ponder the Bits (2017) *OSX/Mac memory acquisition and analysis using OSXPMem and Volatility*. Available at: https://ponderthebits.com/2017/02/osx-mac-memory-acquisition-and-analysis-using-osxpmem-and-volatility/
SalvationDATA (2023) *Forensic timeline analysis*. Available at: https://www.salvationdata.com/knowledge/forensic-timeline/
SentinelOne (2024) *Digital forensics: Definition and best practices*. Available at: https://www.sentinelone.com/cybersecurity-101/cybersecurity/digital-forensics/
Sleuth Kit (2001) *The Sleuth Kit (TSK): Open-source digital forensics tools*. Available at: https://www.sleuthkit.org/sleuthkit
Stack Overflow (2014) *Issues compiling Android kernel (cross-compiling) with LiME*. Available at: https://stackoverflow.com/questions/26720566/issues-compiling-android-kernel-cross-compiling-with-lime
Sudhakaran, S. (2022) *Performing memory forensics for object recovery from Android application memory*.
Sudhakaran, S., Ali-Gombe, A., Case, A. and Richard III, G.G. (2022) 'Evaluating the reliability of Android userland memory forensics', *International Conference on Cyber Warfare and Security*, 17(1), pp. 423–432.
Sudhakaran, S., Ali-Gombe, A., Orgah, A., Case, A. and Richard III, G.G. (2020) 'AmpleDroid: Recovering large object files from application memory', *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6.
Sudhakaran, S., Nur, N. and Zarraa, A. (2024) 'Navigating the challenges: Understanding anxiety among international graduate students in the United States', *INTED2024 Proceedings*, pp. 7664–7672.
TechFusion (2024) *Chain of custody in digital forensics*. Available at: https://techfusion.com/chain-of-custody-in-digital-forensics/
Volatility Foundation (2014) *Volatility: Advanced memory forensics framework*. Available at: https://www.volatilityfoundation.org

Regulatory Challenges in Maritime Cybersecurity: Evidence from Expert Workshops

Bilge Karabacak, Kasey Miller, Ulku Clark, Jeff Cummings, Edwin Garces and Geoff Stoker
Congdon School of Supply Chain, Business Analytics, & Information Systems, Cameron School of Business, Wilmington, NC, USA

karabacakb@uncw.edu
millerkc@uncw.edu
clarku@uncw.edu
cummingsj@uncw.edu
garcese@uncw.edu
stokerg@uncw.edu

Abstract: The maritime transportation sector is crucial for nations worldwide, as the Maritime Transportation System (MTS) facilitates over 90% of international trade through an extensive network of ships and ports. As digitized maritime operations become increasingly prevalent, advanced technologies are being integrated into these systems. MTS's critical importance, along with its dependence on interconnected systems and new digital technologies, renders it vulnerable to cyberattacks at ports and other key points. Consequently, implementing cybersecurity regulations in the maritime sector is essential to ensure operational safety and security. However, organizations within this sector encounter significant challenges in adopting and adapting to these regulations, highlighting the need for robust, clear, and enforceable standards. Previous studies have identified various gaps and challenges in maritime cybersecurity regulations, including fragmented, outdated, and inconsistent enforcement. To address these issues, this paper employs a qualitative analysis through expert workshops to evaluate whether the findings effectively reflect real-world experiences with operational and regulatory applications. The study aims to identify and analyze gaps and challenges in existing regulations, policy standards, and frameworks for maritime cybersecurity.

Keywords: Maritime cybersecurity, Regulations, Standards, Frameworks, Qualitative data analysis, Expert elicitation, Cyber resilience

1. Introduction

Maritime logistics is a crucial component of global business and freight, encompassing over 90% of world trade and more than 50,000 ships (Drummond and Machado, 2021). Consequently, the integration of emerging maritime technologies has led to a surge in novel cybersecurity threats. The accelerated digitalization of systems has not been met with proper and prompt debugging, release, and regulation (Hopcraft and Martin, 2018; Karim, 2022; Dimakopoulou and Rantos, 2024). Regulations, standards, and frameworks do not always align with real-world practice, causing contradictory diagrams, procedures, and limited integration among voluntary frameworks.

However, regulations, standards, and frameworks are crucial for preventing cybersecurity attacks. They serve as primary sources for security requirements and controls, helping organizations establish robust defenses, protect data, and mitigate risks. Adherence to these requirements supports consistent security practices and builds trust among partners, customers, and stakeholders (Taherdoost, 2022). Furthermore, mandatory and optional frameworks provide essential guidance to support national security and raise awareness of cybersecurity risks (Djebbar and Nordstrom, 2023; Amin, 2024; Folorunso, Mohammed, et al., 2024; Folorunso, Wada, et al., 2024).

One challenge in maritime cybersecurity is the lack of regulatory specificity for particular fields or sectors. Non-tailored regulations can lead to inefficient operations, non-compliance, reduced control, and increased vulnerability, resulting in data breaches, financial losses, and reputational damage (Olney, 2023; Anchore, 2025).

Each sector operates differently and has unique vulnerabilities; consequently, cybersecurity requirements also differ (Miller, 2024). Regulations therefore need to be tailored to specific domains, particularly as automation and new technologies increase system complexity (Heckman et al., 2017; Dimakopoulou and Rantos, 2024). While compliance with standards remains essential, the development and adaptation of effective standards remain unresolved challenges.

Several authors note that experts face difficulties interpreting fragmented and unevenly applied rules, especially when multiple frameworks coexist across national and international contexts (Hopcraft and Martin, 2018; Drazovich, Brew, and Wetzel, 2021; Melnyk, Drozdov, and Kuznichenko, 2025). This underscores the need for

studies identifying gaps and challenges to support harmonized and well-structured regulatory systems (Hopcraft and Martin, 2018; Karim, 2022).

Collecting input from maritime cybersecurity professionals is critical because it provides practical insights that theory or policy alone cannot capture. These insights highlight problems, gaps, and unmet requirements in maritime transportation system regulations (Park et al., 2023). Accordingly, we organized two Maritime Technology Road Mapping (TRM) workshops in January and March 2025 (Clark et al., 2025), interviewing 38 experts from government, NGOs, and the private sector. This study analyzes regulatory challenges identified in these workshops and categorizes regulatory gaps to inform the development of more robust and standardized frameworks.

2. Literature Review

The literature review is organized into two subsections. In the first subsection, we provide an overview of the current cybersecurity regulations, standards, and frameworks in the maritime sector. In the second subsection, we present experts' perspectives on the need for regulatory improvements.

2.1 Regulations, Standards, and Frameworks in Maritime Cybersecurity

The documentation on maritime cybersecurity can be categorized into international instruments, national and international standards, and specialized assessment frameworks.

For international instruments, the International Maritime Organization (IMO) plays a central role. IMO established Resolution MSC.428(98), mandating the inclusion of cyber risk management in Safety Management Systems (SMS), and issued non-mandatory high-level guidelines (Cyber Risk Management – MSC-FAL.1/Circ. 3). Existing codes include the International Ship and Port Facility Security (ISPS) Code and the ISM Code for safety management (Hopcraft and Martin, 2018; Svilicic et al., 2019; Chupkemi and Mersinas, 2024).

National and international standards include the National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF) version 2.0, a widely recognized resource for cybersecurity management. NIST CSF 2.0 consists of six functions: Govern, Identify, Protect, Detect, Respond, and Recover. Another key source of guidance comes from classification societies and industry organizations such as BIMCO, DNV-GL, the International Association of Classification Societies (IACS), and the American Bureau of Shipping (ABS). IACS provides specific cyber-related requirements through UR E26 (Cyber Resilience of Ships) and UR E27 (Cyber Resilience of Onboard Systems and Equipment). UR E26, adopted in 2022, establishes minimum cyber-resilience requirements at the ship level, while UR E27 defines cybersecurity requirements for onboard systems, including secure development and vendor obligations. Both requirements are enforced for newbuild ships from July 1, 2024 (IACS, 2022b, 2022a).

Specialized assessment frameworks have been evaluated in studies proposing targeted approaches for maritime cybersecurity. For ships, the Maritime Cyber Risk Assessment (MaCRA) model quantifies cyber risks based on system vulnerabilities, ease of exploitation, and attacker rewards (Tam and Jones, 2019). Other proposed frameworks include Cyber Resilience Evaluation models aligned with UR E26, the NIST CSF, the MITRE Cyber Resiliency Engineering Framework, and the Ship Cyber Risk Assessment Framework (Ko, Lee, and Seo, 2025). The Threat-Informed Defense-in-Depth approach combines threat-informed defense and layered security strategies, particularly for autonomous passenger ships (Amro and Gkioulos, 2023; Wu et al., 2023; Dimakopoulou and Rantos, 2024).

While these frameworks focus on vessels, other models address ports and critical infrastructure. These include the Integrated Cyber Risk Assessment (ICSRA), which applies ISO 31000, NIST SP800-30, and NERC CIP to support risk understanding, monitoring, and management (Drummond and Machado, 2021; Gunes, Kayisoglu and Bolat, 2021). Another model, the Maritime Cyber Assessment Organization (MCAO), focuses on port operations and critical infrastructure, evaluating regulatory structures and proposing a holistic public–private assessment approach to establish and enforce cybersecurity standards (Trimble, Monken and Sand, 2017).

2.2 Literature about Experts' Perspectives for Improving Regulations

These studies have expert contribute to the academic dialogue on current regulations focus on non-binding guidelines and the subjective interpretation of instruments, aiming to develop better, more robust, and specialized frameworks (Ogundare and Akinwande, 2021; Karim, 2022). Their aim is to foster the development of more robust and specialized frameworks. A significant body of literature advocates for the creation of a standalone cyber code by the IMO to improve legal binding, harmonization, and enforcement, as well as enhance

clarity and modernization (Hopcraft and Martin, 2018; Ogundare and Akinwande, 2021). At the same time, some studies propose focusing on strengthening the existing frameworks and codes. The suggestions include improving the clarity, highlighting the need to strengthen the ISM Code, enhancing clarity and deterrence, and promoting multi-level cooperation and technical customization, as seen in the case of the IACS UR E26 framework, which requires enhancements to a unique environment for ships (Ogundare and Akinwande, 2021; Karim, 2022; Dimakopoulou and Rantos, 2024).

Simultaneously, some studies suggest different tools for standardizing organizational and methodological measures for training and educating personnel. These approaches focus on reducing human error in cybersecurity by suggesting minimum standards. These studies suggest improving quantitative risk assessment methods by considering both Information Technology (IT) and Operational Technology (OT) systems (Sobiesk, Bennett and Maxwell, 2017; Drazovich, Brew and Wetzel, 2021; Drummond and Machado, 2021; Soner *et al.*, 2024).

The existing literature reveals a gap in systematically incorporating stakeholder expert opinions and objectively validating these perspectives. In this study, a total of 38 experts were identified using a Network-Referral-Based method (own pre-existing network), Chain Referral Sampling, and Bibliographic and patent social network analysis. The panels of experts were adequately balanced among representatives from industry, academia, government, and NGOs.

3. Methodology

This study employs a qualitative approach, drawing on experts with experience in maritime cybersecurity and a solid understanding of both operational and regulatory issues. The regulatory frameworks mentioned previously were discussed at length during the maritime cybersecurity TRM sessions. The TRM session was attended by 38 experts from institutions across North America and Europe, representing academia, government, and private industry. The TRM project commenced in October 2024 with an initial stakeholder meeting that laid the groundwork for the subsequent two workshops, which systematically addressed all six NIST Cybersecurity Framework functions (Govern, Identify, Protect, Detect, Respond, and Recover). The

TRM was structured into four "swim lanes": Drivers, Capability Gaps, Technology Characteristics, and R&D programs. For this study, information on Drivers and Capability Gaps was used, as they are directly related to the identification of challenges and gaps. In this context, Drivers are considered as factors associated with organizational changes or decisions, including upcoming regulations and standards (Hillegas-Elting *et al.*, 2015).

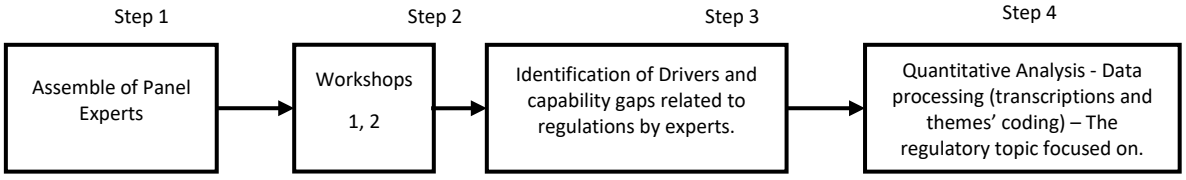


Figure 1: Research Methodology Flow

The process of selecting experts involves identification, contact, engagement, and panel formation. Experts were identified using the Network-Referral-Based method, Chain Referral Sampling, and Bibliographic and patent social network analysis, in accordance with the inclusion criteria of relevant knowledge and experience in maritime cybersecurity, as well as reputation in industry, government, and academia**.** To minimize bias, the selection and invitation of these experts were balanced within each NIST function by expertise type, organizational role, and regional location.

The study utilizes information obtained from two online workshops conducted to develop a technology roadmap. The workshops counted on the participation of 38 experts. These two workshops were organized with distinct objectives, with discussions focusing on technological and regulatory challenges and solutions in maritime cybersecurity. Workshop 1 focused on identifying drivers and capability gaps in the maritime cybersecurity field, with participation from stakeholders, managers, decision-makers, and directors. As mentioned above, the information obtained from this workshop served as a data source for the present study, as it is directly relevant to identifying regulatory challenges. The second workshop focused on technological availability and on new technologies that need to be developed to address the gaps and drivers. The third step involves analytical and qualitative analysis, which includes identifying regulatory challenges and phrases and paragraphs directly related to regulatory elements, thereby coding and grouping them by theme. The data

corresponds to written and audio files. These themes were classified into three categories (governance, people, and technology) based on the type of challenges.

4. Results and Data Analysis

The results of the qualitative analysis from the expert discussions during the two workshops are included in Tables 1, 2, and 3. The process includes three main steps: data transcription, data preparation and standardization, and identification of qualitative components directly related to regulatory topics. These components were organized according to the maritime cybersecurity TRM (Clark *et al.*, 2025) areas and NIST CSF functions. The identified discussions about regulatory challenges were grouped, analyzed by themes, and assigned to their corresponding category. For each theme, similar ideas were merged, and related regulatory references were noted. The analysis also includes identifying the corresponding NIST CSF functions based on panel discussions. Discussions on regulatory challenges were grouped by theme, analyzed, and assigned to their corresponding category. For each theme, similar ideas and opinions were merged, and related regulatory references were noted. Finally, the analysis linked these themes to the appropriate NIST CSF functions based on the panel discussions.

Each table shows the identified regulatory challenges and gaps, categorized into three groups: Governance & Strategy (see Table 1), People & Process (see Table 2), and Technology & Implementation (see Table 3). Each category contains subcategories that explain the challenges and needs associated with it, the type of norms it encompasses, the kinds of security policies it belongs to, and the NIST CSF function associated with the topic. This last aspect was included to maintain the data logically categorized within the context in which the technology roadmap was developed, considering the NIST CSF functions as the roadmap areas.

Table 1: Governance & Strategy (Category 1)

Subcategory	Challenges and Needs	Type of Regulation or Standard, or Framework Referred to	NIST Function
Budgeting and Resource Allocation Gaps	Budget constraints and inconsistent funding impact upgrades and system innovation. The effects include under-prioritized cybersecurity efforts. C-suite executives often have only the minimum required for compliance, related to the high costs of complex systems (integration of IT/OT systems).	IMO requirement (risk management) Regulations (general) R&D Funding	Identify Govern Detect Protect
Regulatory Structure and Clarity (Jurisdiction/Vision) Gaps	The regulatory governance structure and jurisdiction are fragmented and unclear. The results include having undefined cybersecurity enforcement responsibilities. The causes of these issues are related to the lack of a single authority, inconsistent governance, and challenges in implementing a unified risk strategy due to disagreements on core definitions.	ISO Risk Management Strategy	Identify Govern
Standards Development & Maintenance Gaps	Current IACS standards, such as E26/E27, are not effectively enforced on in-service vessels. Additionally, unclear rules and slow regulatory updates affect the responses to attacks. There is a lack of a unified certification framework for IT/OT/IoT devices, and integrating cybersecurity into the ISPS Code is a slow process. It is needed more dynamic standards' ecosystems, as well as more specific and authoritative standards.	IACS Standards (E26, E27) Certification and Accreditation Framework/Authority Regulations and Standards International Ship and Port Facility Security (ISPS) Code framework	Protect Detect Identify
Third-Party/Vendor Risk Management Gaps	External partners or vendors involved in maritime operations often lack awareness of regulatory requirements and formal certification programs, resulting in compliance and security gaps. The decision-making process is weak, where vendors are rarely included in the plans. These plans do not include all the stakeholders, and instead, they are driven by other factors, such as required insurance. The strategic plans should include vendors, regulators, and operators.	Certification Programs Risk Management Platforms Cyber Security Insurance Policies - Institute Cyber Attack Exclusion Clause (CL 380)	Protect Identify Recover Govern

Subcategory	Challenges and Needs	Type of Regulation or Standard, or Framework Referred to	NIST Function
Weak Cybersecurity Governance and Strategy Gaps	Executive leadership often does not treat cybersecurity as a critical business priority. As a result, organizations allocate insufficient funds and hire few personnel. The effects of these practices include excessive pressure on personnel, which can lead to an increase in human errors and the implementation of frameworks poorly. To solve these problems, it is necessary to have proper funding, clear strategic guidance, and ensure that cybersecurity measures are coordinated among all the actors.	Cyber Risk Management Risk Management Framework	Identify Govern

Table 2: People & Process (Category 2)

Subcategory	Challenges and Needs	Type of Regulation or Standard, or Framework Referred to	NIST Function
Compliance Culture and Minimum Requirements Gap	Maritime organizations often undervalue cybersecurity, resulting in compliance with the minimum mandatory requirements, which weakens the system's capacity for resilience.	NIST 2.0 International Association of Classification Societies (IACS) requirements Mandatory Certification Class Certification (Class Societies) IMO requirements and guidelines Vetting Requirements (Ship Inspection Report Programme - SIRE, Chemical Distribution Institute - CDI ship, RightShip) Compliance Regulatory Audits	Identify Protect Detect Govern Recover
Expertise, Training, and Awareness Gap	Maritime organizations frequently face a shortage of personnel with advanced cybersecurity expertise in regulatory awareness, particularly among inspectors and senior leaders. These inconsistent onboarding issues make it difficult for organizations to implement cybersecurity strategies effectively. In addition to having an adequate number of personnel, it is necessary to implement role-specific training programs, standardized training, and have a solid understanding of evolving requirements.	Certification and Training Programs IMO standards and regulations Cybersecurity readiness assessment	Detect Respond Protect Identify Recover Govern
Incident Response and Business Continuity Gap	The regulatory entities offer limited guidance and inconsistent reporting procedures, as maritime organizations often lack comprehensive cybersecurity business continuity plans, standardized drills, and coordinated incident response procedures. Therefore, a unified framework and proactive preparedness are crucial to enhancing resilience across the sector.	Incident Response - IR Planning Cybersecurity Drills and Simulations Safety Management System (SMS), Ship Security Plan (SSP), Ship Security Assessment (SSA) Business Continuity Plans (BCPs)	Respond
Procedure and Documentation Gap	Ports and maritime organizations often lack standardized incident procedures. Organizations generally have inconsistent reporting and outdated recovery guidelines that are poorly integrated into formal documentation. There are inadequate audits, shared policy templates, and uniform documentation practices, which negatively impact strategic planning and communication. A centralized, updated framework is needed to support consistent and effective cyber incident response.	ISM code SMS (Safety Management System) Post Response Reporting Procedures Incident procedures Audits, Procedures Cyber Policy Templates Recovery Guidelines	Respond Identify Recover

Table 3: Technology & Implementation (Category 3)

Subcategory	Challenges or Needs Related to the Standard/Regulation/Frameworks	Type of Regulation or Standard, or Framework Referred to	NIST Function
Asset Management and Inventory Gap	Maritime operations often lack complete asset documentation, onboard cybersecurity ownership, and limited visibility into system vulnerabilities and threats. This implies having untrained personnel for managing critical systems and poorly tracked supply chain risks (embedded modems in port equipment).	Asset Documentation Cybersecurity Ownership IT/OT Asset Identification	Identify
Design, Technology, and Legacy System Integration Gap	Legacy ship systems were not designed for modern cybersecurity technologies, facing challenges in integrating new technologies (e.g., Zero Trust or SIEM). This aspect is associated with system designs and updates. The organizations also lack a consistent evaluation of IoT frameworks and flexible R&D alignment with standards, despite guidance such as IACS UR E26. Therefore, a strategic approach is warranted to bridge design limitations and changes in technological dynamics.	Zero Trust model Security Information and Event Management - SIEM (frameworks/tools) NIST SP800-213 and 213A (IoT frameworks) International Association of Classification Societies (IACS) Unified Requirement UR E26 guidelines Network Standards Series of international standards for the cybersecurity of industrial automation and control systems, IEC 62443 (International Electrotechnical Commission) NIST CSF International Organization for Standardization - ISO 27001 standards	Respond Detect Protect Identify
Implementation & Enforcement Gap	Maritime cybersecurity remains vulnerable to weak enforcement, high compliance costs (meeting cybersecurity requirements often involves expensive upgrades, audits, and certifications), and slow integration into formal systems such as Safety Management Systems (SMS) or International Safety Management (ISM). Stronger regulatory mechanisms, more precise guidance, and streamlined implementation are needed to close the gap between policy and practice.	International Safety Management - ISM code SMS (Safety Management System) Procedures Ship Security Plan - SSP Ship Security Assessment - SSA Class approval IACS Standards (E26, E27) Audits, Procedures Vetting programs	Identify Detect Protect Respond Recover Govern
Security Monitoring and Detection Gap	Older maritime systems lack event monitoring and standardized tools (i.e., SIEM) to detect anomalies in real time. Furthermore, there exists poor threat intelligence sharing across fleets and organizations, reliance on outdated software, and vendor reluctance to report vulnerabilities to protect their reputations. Clear guidelines, built-in security design, and maritime-specific monitoring frameworks are warranted.	Security Event Monitoring Security Information and Event Management - SIEM Security Operations Center - SOC Regulatory Guidelines Mandatory Reporting Requirements	Detect Identify Govern
Vulnerability and Patch Management Gap	Vulnerability management in maritime systems is influenced by legacy OT environments, a lack of built-in software security, and vendor reluctance to disclose flaws due to certification risks. The absence of proactive strengthening strategies and industry-wide debate on secure software design due to the risk that software could be developed without secure coding	Class Certification (Class Societies) Software Quality Requirements	Detect Protect

Subcategory	Challenges or Needs Related to the Standard/Regulation/Frameworks	Type of Regulation or Standard, or Framework Referred to	NIST Function
	practices, leaving it exposed to common vulnerabilities.		

5. Analysis

The results suggest that maritime cybersecurity challenges are linked to fragmented regulatory structures, because maritime cybersecurity rules come from multiple organizations (IMO, flag states, classification societies) and they are not harmonized. Therefore, although regulations exist, enforcement varies by region and authority, resulting in inconsistent enforcement, including minimal compliance and a tendency to maintain outdated infrastructure.

The policies most commonly referenced throughout the study are consistent with established standards, procedures, and frameworks, highlighting a necessity for more structured and adaptable controls. Each subcategory within the three primary categories corresponds to various NIST Cybersecurity Framework (CSF) functions; in fact, all CSF functions are accounted for in this mapping. This alignment is anticipated, as the NIST functions encompass all stages of an organization's cybersecurity lifecycle.

The challenges found within category 1, Governance and Strategy (G), encompass a range of issues, from economic and financial concerns to the structure, development, update, and governance of norms aw well as problems stemming from fragmented organizational structures. Budgeting and resource allocation gaps include compliance requirements and the under-prioritization of cybersecurity efforts. The regulatory structure, with its lack of clarity, is associated with fragmented governance and unclear lines of responsibility. These deficiencies affect the development and maintenance of standards such as E26/E27, which are perceived as insufficiently enforced. Overall, governance and strategy appear fragmented due to limited executive support, the absence of a unified strategy, and the lack of a cohesive certification framework. Weak governance is a cross-cutting issue affecting all topic categories and the risk management process, ultimately hindering the effective implementation of norms.

In the People and Process (P) category, challenges relate to human factors and organizational practices, including gaps in organizational procedures and competencies, operational responses, and documentation. While the causes of this vary, a common theme is that organizations often settle for minimal compliance, which limits the adoption of new protective technologies and adequate personnel training. Other contributing factors include a lack of cyber expertise, standardized training, and regulatory awareness across all organizational levels. These human-centered challenges are linked to a limited understanding of diverse requirements and insufficient onboarding or in-service training. Operational response and documentation issues further indicate the absence of consistent cybersecurity business plans and coordinated response mechanisms across maritime organizations. In this context, establishing standardized incident procedures and maintaining up-to-date policy and guideline libraries is essential.

The last category, Technology & Implementation (T), focuses on technical systems and their deployment. The subcategories address challenges in system infrastructure, design flaws, operational monitoring, and enforcement gaps. A sequential set of issues emerges across asset management, system design, legacy system integration, and vulnerability or patch management. Weak assessments in these areas hinder the effective integration of modern systems into legacy or complex infrastructures, resulting in unmanaged vulnerabilities. Operational monitoring and enforcement gaps include shortcomings in implementing and enforcing existing regulations, particularly those related to the ISM Code and IACS standards. Additionally, security monitoring and detection systems remain inadequate because they rely on outdated tools (e.g., GPS/AIS spoofing detection) and suffer from insufficient information-sharing practices.

6. Conclusion

This paper identified, categorized, and analyzed challenges and gaps within the maritime cybersecurity regulatory framework. Using a qualitative approach, the study draws on insights from 38 subject-matter experts. During two online workshops, experts discussed drivers and capability gaps related to maritime cybersecurity, including regulatory issues. Identified challenges were classified into three categories: governance and strategy, people and processes, and technology and implementation.

The study evaluated Governance & Strategy (G), People & Process (P), and Technology & Implementation (T), encompassing fourteen subcategories addressing key aspects of maritime cybersecurity. Analysis of these categories and recurring challenges reveals systemic gaps, particularly a fragmented regulatory framework combined with weak enforcement, minimal compliance, and unclear norms. This culture of minimal compliance negatively affects operational security. In addition, shortcomings in human capital, expertise development, training, and awareness are linked to insufficient standardized procedures and documentation.

Human-related challenges are critical to fostering stronger compliance cultures, ensuring sufficient expertise, and maintaining properly documented systems. Technological issues are closely connected, as many systems rely on outdated installations that cannot readily integrate modern cybersecurity protections. Legacy technologies pose significant risks because they were not designed to meet contemporary cybersecurity requirements. Even when upgrades occur, poor planning and inadequate standards often lead to incomplete or ineffective controls. Weak assessment management of legacy systems further limits integration of modern security tools, exposing management gaps and undermining security monitoring and enforcement. As a result, regulatory mechanisms fall short of supporting effective operational protection.

This analysis is based on expert perspectives, which introduces subjectivity but is essential for understanding regulatory application in practice. Future studies incorporating academic literature could strengthen the analysis and validate findings across complementary data sources.

References

Amin, M. (2024) "The Importance of Cybersecurity and Protecting of Digital Assets and Understanding the Role of Cybersecurity Laws in Safeguarding Digital Assets," *Indian Journal of Public Administration*, 70(3), pp. 493–501. Available at: https://doi.org/10.1177/00195561241271520.

Amro, A. and Gkioulos, V. (2023) "Cyber risk management for autonomous passenger ships using threat-informed defense-in-depth," *International Journal of Information Security*, 22(1), pp. 249–288. Available at: https://doi.org/10.1007/s10207-022-00638-y.

Anchore (2025) *Your Guide to Cybersecurity Compliance, from Federal Policy to Industry Standards*. Available at: https://anchore.com/compliance/ (Accessed: November 9, 2025).

Chupkemi, D.C. and Mersinas, K. (2024) "Challenges in Maritime Cybersecurity Training and Compliance," *Journal of Marine Science and Engineering*, 12(10). Available at: https://doi.org/10.3390/jmse12101844.

Clark, U. *et al.* (2025) "Maritime Transportation System (MTS) Cybersecurity Technology Roadmap (TRM) Version 1." Available at: https://hdl.handle.net/20.500.14481/1433 (Accessed: November 16, 2025).

Dimakopoulou, A. and Rantos, K. (2024) "Comprehensive Analysis of Maritime Cybersecurity Landscape Based on the NIST CSF v2.0," *Journal of Marine Science and Engineering*, 12(6). Available at: https://doi.org/10.3390/jmse12060919.

Djebbar, F. and Nordstrom, K. (2023) "A Comparative Analysis of Industrial Cybersecurity Standards," *IEEE Access*, 11, pp. 85315–85332. Available at: https://doi.org/10.1109/ACCESS.2023.3303205.

Drazovich, L., Brew, L. and Wetzel, S. (2021) "Advancing the state of maritime cybersecurity guidelines to improve the resilience of the maritime transportation system," in *Proceedings of the 2021 IEEE International Conference on Cyber Security and Resilience, CSR 2021*. Institute of Electrical and Electronics Engineers Inc., pp. 503–509. Available at: https://doi.org/10.1109/CSR51186.2021.9527922.

Drummond, B.M. and Machado, R.C.S. (2021) "Cyber Security Risk Management for Ports - A Systematic Literature Review," in *2021 IEEE International Workshop on Metrology for the Sea: Learning to Measure Sea Health Parameters, MetroSea 2021 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., pp. 406–411. Available at: https://doi.org/10.1109/MetroSea52177.2021.9611569.

Folorunso, A., Wada, I., *et al.* (2024) "Security compliance and its implication for cybersecurity," *https://wjarr.com/sites/default/files/WJARR-2024-3170.pdf*, 24(1), pp. 2105–2121. Available at: https://doi.org/10.30574/WJARR.2024.24.1.3170.

Folorunso, A., Mohammed, V., *et al.* (2024) "The impact of ISO security standards on enhancing cybersecurity posture in organizations," *https://wjarr.com/sites/default/files/WJARR-2024-3169.pdf*, 24(1), pp. 2582–2595. Available at: https://doi.org/10.30574/WJARR.2024.24.1.3169.

Gunes, B., Kayisoglu, G. and Bolat, P. (2021) "Cyber security risk assessment for seaports: A case study of a container port," *Computers & Security*, 103, p. 102196. Available at: https://doi.org/10.1016/J.COSE.2021.102196.

Heckman, M.R. *et al.* (2017) "Chapter 32: Toward a Maritime Cyber Security Compliance Regime," in J. Direnzo III, N.K. Drumhiller, and F.S. Roberts (eds.) *Issues in Maritime Cyber Security*. Washington DC: Westphalia Press, pp. 543–568.

Hillegas-Elting, J. V. *et al.* (2015) "Opening the door to breakthroughs that address strategic organizational needs: Applying technology roadmapping tools and techniques at an electric utility," *Portland International Conference on Management of Engineering and Technology*, 2015-September, pp. 2564–2573. Available at: https://doi.org/10.1109/PICMET.2015.7273034.

Hopcraft, R. and Martin, K.M. (2018) "Effective maritime cybersecurity regulation–the case for a cyber code," *Journal of the Indian Ocean Region*, 14(3), pp. 354–366. Available at: https://doi.org/10.1080/19480881.2018.1519056.

International Association of Classification Societies IACS (2022a) *UR E26 REV1 -2023 Cyber Resilience of Ships*, *E26*. E26. Available at: https://iacs.org.uk/resolutions/unified-requirements/ur-e/ur-e26-new (Accessed: November 22, 2025).

International Association of Classification Societies IACS (2022b) *UR E27 Rev1 2023 - Cyber Resilience of onboard Systems and Equipment*, *E27*. 2022. Available at: https://iacs.org.uk/resolutions/unified-requirements/ur-e/ur-e27-rev1 (Accessed: November 22, 2025).

Karim, M.S. (2022) "Maritime cybersecurity and the IMO legal instruments: Sluggish response to an escalating threat?," *Marine Policy*, 143. Available at: https://doi.org/10.1016/j.marpol.2022.105138.

Ko, A., Lee, J.H. and Seo, J.T. (2025) "KPI-Based Evaluation Framework for Cyber Resilience of Ships," *IEEE Access*, 13, pp. 64226–64245. Available at: https://doi.org/10.1109/ACCESS.2025.3550501.

Melnyk, O., Drozdov, O. and Kuznichenko, S. (2025) "Cybersecurity in Maritime Transport: An International Perspective on Regulatory Frameworks and Countermeasures," *Lex Portus*, 11(1), pp. 7–19. Available at: https://doi.org/10.62821/lp11101.

Miller, A. (2024) *Tailoring Cybersecurity Measures for Different Industries - AgileBlue*. Available at: https://agileblue.com/resource/tailoring-cybersecurity-measures-for-different-industries-2/ (Accessed: November 9, 2025).

Ogundare, B. and Akinwande, G. (2021) "International Maritime Organisation Framework on Cyber Risk Management-a Case for a Comprehensive Legal Framework." Available at: www.imo.org/en/OurWork/Security/Guide_to_Maritime_Security/Pages/Cyber-se-.

Olney, M. (2023) *What are the benefits of cyber security compliance and the consequences of non-compliance?* Available at: https://insights.integrity360.com/what-are-the-benefits-of-cyber-security-compliance-and-the-consequences-of-non-compliance (Accessed: November 9, 2025).

Park, C. *et al.* (2023) "A BN driven FMEA approach to assess maritime cybersecurity risks," *Ocean & Coastal Management*, 235, p. 106480. Available at: https://doi.org/10.1016/J.OCECOAMAN.2023.106480.

Sobiesk, Edward., Bennett, Daniel. and Maxwell, Paul. (2017) "2A Framework for Cybersecurity Assessments of Critical Port Infrastructure," in *2017 IEEE International Conference on Cyber Conflict (CyCon U.S.) : proceedings : 7-8 November 2017, Washington, DC, USA*. IEEE, p. IEEE.

Soner, O. *et al.* (2024) "Risk sensitivity analysis of AIS cyber security through maritime cyber regulatory frameworks," *Applied Ocean Research*, 142. Available at: https://doi.org/10.1016/j.apor.2023.103855.

Svilicic, B. *et al.* (2019) "Assessing ship cyber risks: a framework and case study of ECDIS security," *WMU Journal of Maritime Affairs*, 18(3), pp. 509–520. Available at: https://doi.org/10.1007/s13437-019-00183-x.

Taherdoost, H. (2022) "Understanding Cybersecurity Frameworks and Information Security Standards—A Review and Comprehensive Overview," *Electronics 2022, Vol. 11,* 11(14). Available at: https://doi.org/10.3390/ELECTRONICS11142181.

Tam, K. and Jones, K. (2019) "MaCRA: a model-based framework for maritime cyber-risk assessment," *WMU Journal of Maritime Affairs*, 18(1), pp. 129–163. Available at: https://doi.org/10.1007/s13437-019-00162-2.

Trimble, D., Monken, J. and Sand, A.F.L. (2017) "A Framework for Cybersecurity Assessments of Critical Port Infrastructure," in Edward. Sobiesk, Daniel. Bennett, and Paul. Maxwell (eds.) *2017 IEEE International Conference on Cyber Conflict (CyCon U.S.) : proceedings : 7-8 November 2017, Washington, DC, USA*. New York City, USA: IEEE.

Wu, Z. *et al.* (2023) "An attack-aware shipping enterprise cybersecurity framework based on deep learning," in *Proceedings - 2023 11th International Conference on Information Systems and Computing Technology, ISCTech 2023*. Institute of Electrical and Electronics Engineers Inc., pp. 115–119. Available at: https://doi.org/10.1109/ISCTech60480.2023.00028

Migrating Time and Security-Critical PKIs to Post-Quantum Cryptography: SWIM and C-ITS

Anni Karinsalo, Sara Nikula and Sami Lehtonen
VTT Technical Research Centre of Finland, Oulu/Espoo, Finland

anni.karinsalo@vtt.fi
sara.nikula@vtt.fi
sami.lehtonen@vtt.fi

Abstract: Public Key Infrastructures (PKIs) are foundational to time- and security-critical systems such as air traffic management and cooperative intelligent transport systems (C-ITS). These PKIs rely almost exclusively on classical public-key cryptography, which will become vulnerable once cryptographically relevant quantum computers emerge. Migrating such systems to post-quantum cryptography (PQC) is therefore necessary but non-trivial, as these environments impose strict constraints on latency, bandwidth, interoperability, long system lifetimes, and regulatory compliance. This paper presents a comparative, constraint-driven analysis of PQC migration for time- and security-critical PKIs. Focusing on System-Wide Information Management (SWIM) and C-ITS, we examine how PQC signature and key sizes, verification costs, and certificate structures affect real-time communication, certificate validation, and session establishment. In particular, we analyze the impact of certificate size growth and verification latency on time-critical messaging using published benchmarks and protocol specifications. Rather than proposing new cryptographic primitives or implementations, this work synthesizes existing benchmarks, standardization documents, and protocol specifications to identify feasibility limits, migration risks, and design trade-offs. The analysis shows that certificate chain length and signature overhead can dominate session establishment time in short communication windows, particularly in C-ITS environments, even when individual cryptographic operations remain computationally efficient. We further discuss the operational risks introduced by hybrid cryptographic deployments, including increased system complexity, negotiation failures, and insecure fallback behavior. In addition, we highlight how long system lifetimes and slow standardization cycles in safety-critical sectors complicate timely cryptographic transitions. The results indicate that migration feasibility is often determined by system-level constraints, such as certificate handling, protocol overhead, interoperability requirements, and regulatory alignment, rather than by the performance of individual PQC algorithms alone. Based on this analysis, we present a benchmarking-based migration framework tailored to critical PKIs, highlighting where hybrid cryptographic approaches are unavoidable, where they introduce new risks, and which classes of PQC algorithms are conditionally viable under strict timing and bandwidth constraints. The paper concludes with concrete recommendations for system designers and policymakers to support crypto-agile PQC migration without compromising operational safety.

Keywords: Post-Quantum cryptography, Public key infrastructure, PQC migration, Cooperative intelligent transport systems, System wide information management

1. Introduction

Public Key Infrastructures (PKIs) form the backbone of secure digital communication by enabling authentication, integrity, and trust establishment between distributed entities. In time- and security-critical systems, such as air traffic management, cooperative intelligent transport systems (C-ITS), and other cyber-physical infrastructures, PKIs are not merely supporting components but safety-relevant mechanisms. In these environments, failures or delays in cryptographic operations can directly affect operational continuity and, in extreme cases, human safety.

Most currently deployed PKIs rely on classical public-key cryptographic algorithms, primarily RSA (Moriarty et al., 2016) and elliptic curve cryptography (ECC) (Miller, 1985, Koblitz, 1987). These algorithms are widely considered vulnerable to sufficiently powerful quantum computers due to Shor's algorithm. While cryptographically relevant quantum computers (CRQC) do not yet exist, the long operational lifetimes of critical systems, combined with the risk of "harvest now, decrypt later" attacks, make proactive migration planning unavoidable. Consequently, post-quantum cryptography (PQC) has emerged as a strategic requirement rather than a theoretical concern.

Migrating PKIs to PQC is significantly more complex in time- and security-critical systems than in conventional IT environments. Such systems impose strict constraints on latency, bandwidth, computational resources, interoperability, and regulatory compliance. PQC algorithms typically introduce substantially larger keys and signatures, which can stress communication channels and certificate validation mechanisms. In addition, critical PKIs are governed by international standards and certification processes that evolve more slowly than cryptographic research, further constraining feasible migration paths and limiting near-term deployment options.

1.1 Scope of This Paper

This paper addresses PQC migration from a systems and PKI-engineering perspective rather than from an algorithm-design or implementation standpoint. We focus on authenticated communication, certificate chain validation, and session establishment, as these are the dominant contributors to cryptographic overhead in PKI-based communication. The paper does not present new cryptographic constructions or experimental implementations; instead, it synthesizes published benchmarks, standardization documents, and protocol specifications to analyse feasibility limits and design trade-offs. We have concentrated specifically on session establishment and certificate chain validation. These two are the main factors in communication between end users and the main contributor in the PKI-based overhead. This work assumes the widely accepted cryptanalytic threat model used by NIST and ETSI, rather than attempting to validate quantum feasibility independently.

Our analysis handles two representative case studies: System-Wide Information Management (SWIM) (ICAO, 2018), which contains emerging aviation data exchange architectures, and Cooperative Intelligent Transport Systems (C-ITS), which enable real-time vehicle-to-everything (V2X) communication. These case studies are used to illustrate how PQC-related performance characteristics interact with real-world timing, bandwidth, and interoperability constraints. Based on this analysis, we derive a migration framework for time- and security-critical PKIs and identify conditions under which hybrid cryptographic approaches are unavoidable, as well as the risks they introduce.

The remainder of this paper is structured as follows. Section 2 provides background on PQC, SWIM, and C-ITS. Section 3 examines these systems as case-study PKIs, focusing on hybrid certificates, algorithm applicability, and certificate lifecycle considerations. Section 4 analyses the main challenges and constraints of PQC migration in time- and security-critical PKIs. Section 5 presents a benchmarking framework for evaluating PQC feasibility under strict performance constraints. Section 6 discusses design implications and migration guidance for system designers and policymakers, followed by conclusions and directions for future work.

Please note: we use the term ‘critical PKI’ as a synonym for time- and security-critical for simplicity.

2. Background and Scope

Ensuring the integrity and origin of transmitted information is traditionally achieved using signature algorithms based on the discrete logarithm problem or the factorization of large numbers. Since 2016, NIST has led the standardization of PQC algorithms. In 2024, NIST published its first PQC standards for digital signatures and key establishment (NIST, 2024a; NIST, 2024b; NIST 2024c). NIST disallows asymmetric cryptography vulnerable to quantum computers by year 2035 (NIST, 2024d).

PKI is a widely adopted framework for secure digital communication and identity verification. In brief, PKI enables entities to authenticate each other and ensure data integrity using asymmetric cryptography, trusted certificate authorities and certificate status queries.

2.1 System-Wide Information Management (SWIM) in Aviation PKIs

Air Traffic Management (ATM) systems rely on short-lived, time-critical communications with strict security and latency requirements.

SWIM is a framework that integrates global real-time data exchange between various aviation systems using common standards, infrastructure, and shared air traffic data. Reference implementations are currently undergoing testing with the aim of future deployment for ATM data management and planning. SWIM is primarily developed under EUROCONTROL in Europe. However, it also plays a significant role in the United States, where joint U.S.-EU strategies are being developed to ensure global harmonization and interoperability, reflecting SWIM’s growing importance in international air traffic management (SESAR-NextGen, 2018). The European Commission (2021) defines the SWIM Yellow Profile as a ground distribution mechanism in which the technical infrastructure can run over any IP-based network.

The European Aviation Common Public Key Infrastructure (EACP) (EUROCONTROL, 2022a) is a European-wide, globally supported shared PKI aimed at enabling secure communication within the European aviation context. EACP provides a trust framework for SWIM communication and data management and will also define rules for local implementations. EACP aligns with European Commission’s mandate (European Commission, 2022) for a common aviation PKI. The European ATM Master Plan (European Commission, 2021) serves as the implementation roadmap for applying EACP to SWIM. Within EACP, civil and military stakeholders may either

operate a local PKI in which they act as a certification authority, use the central EACP CA, or adopt a hybrid approach.

EACP uses the X.509 v3 (ITU, 2000) certificate format, and current CAs rely on ECDSA and RSA signature algorithms (EUROCONTROL, 2022b). SWIM documentation assumes that ECC-224 and RSA-2048 remain secure until approximately 2030, while larger key sizes are projected to remain secure until around 2035. Public SWIM documentation does not disclose any concrete strategy for migration to PQC-compliant algorithms.

2.2 Cooperative Intelligent Transport Systems

Critical communication in intelligent transport systems occurs between nodes called Intelligent Transportation System's Stations (ITS-Ss), including vehicles on the road and static roadside communication points. These nodes communicate using an underlying PKI. ETSI standards define the message and certificate formats used by intelligent transport systems. Some of the defined message types are CAM (ETSI, 2019a) and DENM (ETSI, 2019b), which are used to alert other vehicles to noteworthy events. At the time of writing, ETSI TS 103 097 defines the certificate structures used in C-ITS. Regarding signature types, it refers to IEEE 1609.2, which specifies several elliptic curve options. ETSI TS 102 941 defines the communication between vehicles (ITS stations) and PKI entities in the context of secure and privacy-preserving V2X systems. It specifies the trust models, certificate and credential handling, and secure communication interfaces between vehicles and PKI entities. Currently, ETSI's C-ITS standards provide no support for PQC signatures (ETSI 2022, ETSI 2021, IEEE 2023).

3. Case Study PKIs: PQC Applicability to SWIM and C-ITS

In this section, we examine the practical applicability of PQC to SWIM and C-ITS. As publicly accessible, PQC-enabled SWIM implementations were not available at the time of writing, we use C-ITS as the primary empirical reference case and generalize relevant findings to SWIM messaging and PKI validation constraints.

3.1 Hybrid Certificates

The most widely deployed certificate format for PKIs is X.509 v3, which supports only a single subject public key. This design poses a fundamental limitation for combining classical and post-quantum cryptography within a single certificate. While custom extensions can be used to embed additional public keys, such approaches are not standardized and require explicit support from both the issuing authority and the validating peer. To address this limitation, the Internet Engineering Task Force (IETF) has proposed composite signature schemes in which a single signature is formed from multiple algorithms, such as ECDSA and ML-DSA (Ounsworth et al., 2024).

Alternatively, and more commonly, two separate X.509 certificates are issued and combined using certificate chains or alternate trust anchors, as also supported by NIST. This approach requires communication protocols that can negotiate which certificate and signature algorithm to use, such as TLS 1.3 with PQC extensions. However, under current C-ITS standards, a major limitation is that existing certificate profiles do not support the concurrent use of multiple signature algorithms. Given that vehicles may remain in operation for several decades, it is impractical to transition to post-quantum algorithms in a single coordinated update. As a result, simultaneous support for multiple signature types during extended transition periods becomes unavoidable (Yoshizawa and Preneel, 2023). While ETSI (2023) does not recommend hybrid solutions and instead emphasizes crypto-agility and rapid algorithm replacement, this assumption conflicts with the operational realities of long-lived vehicular systems.

Hybrid certificates introduce risks for both the developer and the user. From the developer perspective, hybrid deployments require maintaining two parallel PKI stacks, one based on classical cryptography and one on post-quantum algorithms. This entails operating two certificate hierarchies and implementing negotiation, fallback, and failover logic between cryptographic stacks. In time-critical systems, this additional complexity introduces risks of increased latency, software defects, and misconfiguration. In the worst case, race conditions or logic errors may cause the system to select an insecure fallback algorithm under load or timing pressure, thereby silently degrading the security level of the communication.

From the user perspective, mixed cryptographic environments introduce a risk of compatibility failures and service disruptions. A client device may support only a single cryptographic algorithm, while the server offers a certificate based on a different algorithm. Without standardized support for dual keying or hybrid validation, such mismatches lead to failed connections or ambiguous certificate errors. As a result, essential operations such as authentication, software updates, or safety messaging may fail unexpectedly. In the worst case, users or systems may disable certificate validation or encryption to restore connectivity, thereby exposing the communication to downgrade attacks or complete loss of confidentiality and integrity.

3.2 Algorithm Applicability and Optimization

NIST is evaluating additional PQC signature schemes with improved or alternative security assumptions (Alagic et al., 2024). Table 1 lists the combined lengths of the public key and signature for algorithms that have either already been standardized or selected for the second round of additional signatures. The algorithms are categorized based on the underlying mathematical problem, and if multiple algorithms based on the same problem have been submitted to the competition, their averages are shown.

Table 1: Average total lengths of signature and public key of digital signature algorithms that are already standardized by NIST or chosen for the second round of additional digital signatures

type	bytes
lattice	1998
hash	7888
multivariate	18903
code-based	17537
symmetric	5038
MPC-in-the-head	8165
isogeny	214
average (all PQC)	9935
RSA (classical)	512
ECDSA (classical)	97

Algorithms based on different mathematical problems exhibit widely varying public key sizes, signature lengths, and computational performance. Consequently, replacing classical signature schemes with post-quantum alternatives impacts both the size of transmitted messages and the time required for signing and verification operations. These effects are particularly relevant in vehicular and aviation communication, where message sizes are tightly constrained and authentication must be completed within short communication windows.

The size of a signature impacts the total length of the transmitted messages. The size of certificates is influenced by both the signatures and the public keys, as they must at least include the signature and public key of the certificate authority, as well as the public key of the certificate owner. (Nikula, Halunen, Vallivaara, 2022) constructed DENM messages defined in (ETSI, 2019b) and calculated that the total length of DENM messages was 300 bytes + length of digital signature + length of public key. If the PKI uses intermediate certificate authorities, the messages become even longer, as their signatures and public keys must also be included in the certificates.

Table 2 presents the size of the DENM message calculated as described in Nikula, Halunen, Vallivaara (2022) assuming the use of the specified signature algorithm, along with the verification time promised by the lowest security level of the signature algorithm (Bos et al., 2025, Aardal et al., 2025, Beullens et al., 2024, Aumasson et al., 2022, Alagic et al., 2022) and ECDSA verification time taken from Nikula, Halunen, Vallivaara (2022).

Table 2: Lengths of DENM messages constructed with different PQC digital signature algorithms and their verification times

type	name	DENM length, bytes	verification time, seconds
lattice	FALCON	1799	3.58×10^{-5}
lattice	HAWK	935	6.44×10^{-5}
hash	SPHINCS+	8124	1.5×10^{-3}
isogeny	SQIsign	450	1.5×10^{-3}
multivariate	MAYO	2110	9.06×10^{-5}
code-based	CROSS	13241	4.22×10^{-4}
elliptic curve (classical)	ECDSA	338	7.8×10^{-5}

Using larger signatures and public keys to form certificates introduces a significant performance constraint. Yoshizawa and Preneel (2023) estimate that the effective channel capacity is reduced by approximately half

when even a subset of vehicles transitions to quantum-safe signature algorithms. Consistently with this observation, ETSI (2023) concludes that none of the NIST-standardized signature schemes, ML-DSA (Dilithium), FN-DSA (Falcon), or SLH-DSA (SPHINCS+), are currently suitable for deployment in C-ITS communication. The primary reason is the large size of public keys and signatures, which causes authentication-related data to consume a disproportionate share of the available channel capacity. For example, signatures produced by the lowest security level of SLH-DSA are approximately thirty times larger than those of elliptic curve schemes (NIST, 2024a; NIST, 2024b), making certificate-based authentication impractical under current C-ITS bandwidth constraints.

Communication in aviation can be compared to communication in intelligent transportation systems, as both involve time-critical messaging that may require rapid responses. The signing time is crucial when creating a message, and the verification time is crucial when receiving a message. Sarıbas and Tonyalı (2022) compared the use of classical and quantum-safe algorithms for TLS handshakes. The results indicate that the performance times of quantum-safe signature algorithms are often even better than those of classical digital signatures. Furthermore, Nikula, Halunen, Vallivaara (2022) claim that lattice-based signature options are suitable for C-ITS communication in terms of performance.

In every communication session, both parties must validate each other's certificate chains by verifying each signature and checking the validity of each certificate in the chain. This requires the exchange of several kilobytes of authentication data, even when the actual application payload may be a few hundred bytes. The communication range of a roadside unit is typically 50 – 200m. For a vehicle travelling at 70 km/h, this corresponds to a communication window of approximately 3s at a range of 50m, and even less at larger distances. Communication between two vehicles is typically reliable only when they travel in the same direction; for oncoming vehicles, the available communication window is extremely short. Since communication sessions are initiated by certificate exchange and validation, certificate size has a direct impact on the probability of successful session establishment. If the certificate chain is too long, the handshake process may consume a significant portion of the available communication time. It is therefore not sufficient that cryptographic operations are computationally feasible; they must also complete within the limited communication window.

4. PQC Migration Challenges and Constraints

Migrating critical PKIs to PQC introduces a set of systemic challenges that extend beyond the choice of cryptographic algorithms. In PKI systems, authentication mechanisms are dependent on real-time communication, bandwidth limitations, long system lifetimes, and regulatory constraints. PQC migration is constrained not only by cryptographic performance, but also by interoperability and backward-compatibility requirements, latency budgets, channel capacity, and sector-specific certification processes. These constraints fundamentally limit the feasibility and pace of migration in critical infrastructures.

When digital certificates are updated to be PQC compatible across certification authorities (CAs), registration authorities, key management systems, directory services, and end users, this requires supporting new cryptographic algorithms in certificate issuance processes and modifying validation and revocation mechanisms. The greatest challenge in this migration is that all related systems utilizing the certificates must be interoperable with the new algorithms. Ensuring backwards compatibility during the transition period is crucial for security.

As illustrated by the C-ITS case study in the previous section, many PQC algorithms introduce increased computational overhead and larger key sizes due to performance and latency constraints. This can impact system performance, especially in real-time communication systems like SWIM and C-ITS. The larger keys and signatures, such as those presented in Table 2, affect certificate issuance, storage, and revocation mechanisms, requiring re-evaluation of key management protocols. Moreover, certificate size affects protocol overhead in mechanisms such as TLS, as certificate fragmentation during transmission increases latency and retransmission risk (Kampanakis et al., 2018).

Certificate agility enables flexible issuance, revocation, and updating of certificates, for example through hybrid certificates and certificate inventories. PKI automation supports this process by enabling automated certificate rotation using systems such as the Automatic Certificate Management Environment (ACME) (Barnes et al., 2019). In this model, lifecycle management is delegated to an ACME client, which must itself be secure and support the selected PQC signature algorithms to avoid introducing new operational vulnerabilities.

Even though technical advances may occur rapidly, regulations and standardizations affect the migration to PQC. Critical PKIs are subject to stringent regulatory requirements and their migration to PQC must particularly align

with evolving international standards from organizations such as ICAO (International Civil Aviation Organization), ETSI, and NIST.

5. Benchmarking Framework for Migrating Critical PKIs

PQC causes new performance constraints in systems where low latency, high availability, and high integrity are essential. Depending on the deployment environment, such as embedded and IoT systems, resource efficiency may also be a mandatory requirement. To evaluate the feasibility of PQC in critical PKIs, we propose a benchmarking framework based on the following performance and communication parameters.

Key generation time: the time required to generate public-private key pairs, especially in when using ephemeral keys. Comparing this with existing RSA/ECDSA key generation times helps understand migration usefulness.

Signature generation and verification time: Signature generation is critical for senders (e.g. vehicles broadcasting safety messages). Verification time is often more performance-critical (e.g. C-ITS road-side units verifying thousands of messages per second). PQC offer trade-offs: FALCON is faster at verification, but more complex to implement securely.

Ciphertext and signature sizes: Larger PQC signatures and ciphertexts can overload communication and increase packet fragmentation. This is important in bandwidth-limited or latency-sensitive systems. For example, a typical PQC signature (Dilithium 2 2.4 KB) is orders of magnitude larger than ECDSA (64 bytes).

Memory and computation footprint: analysing RAM and CPU use in cryptographic operations. Embedded devices, like onboard units (OBUs), have strict resource constraints. Computationally intensive PQC algorithms may be impractical without hardware acceleration.

Latency impact on system-level operations: combined latency of cryptographic operations across protocol stacks. In time-critical domains, even small latency increase can lead to delays or lost messages. Benchmarking must reflect end-to-end scenarios such as certificate validation and message authentication in real-time SWIM or V2X exchanges.

Network throughput and scalability by assessing how the increased size of cryptographic algorithms impacts network throughput. Systems should be then stress tested under high load to determine the number of cryptographic operations per second that can be sustained without degradation.

Hardware acceleration potential: Some PQC can benefit significantly from hardware acceleration. Benchmarking should compare software-only with hardware-accelerated implementations if applicable.

Protocol integration overheads: measuring the added complexity and processing delay with PQC integrated into existing protocols. Hybrid schemes cause additional cryptographic load in multi-algorithm handshakes, which should be reflected in benchmarking.

6. Design Implications and Migration Recommendations

While benchmarking frameworks support the technical evaluation of PQC, effective migration in critical PKIs additionally requires coordinated design strategies, crypto-agile system architectures, and operational planning across the entire PKI ecosystem. For example, regarding the previous C-ITS case study, while the effect of overhead and latency can be estimated, the PQC-enabled system must always be tested against a real-life environment. Standards, specifications and other requirements can give estimate values, but the real-life situation is extremely difficult to simulate, since any deviation may cumulate unexpectedly.

Algorithm selection should be guided by system-specific constraints, including latency budgets, computational resources, and key and signature sizes. For example, in latency-sensitive SWIM environments, Falcon may be preferable for verification-heavy workloads, whereas Dilithium may be better suited for balanced signing and verification scenarios. There is no universally applicable algorithm for all critical PKIs, and algorithm choices must therefore be adapted to the operational context.

PQC migration should be implemented in phases, with initial deployment in non-critical system components. Support for classical cryptography should be maintained during transition periods to reduce migration risk and preserve service continuity. Hybrid solutions provide backward compatibility but increase system complexity and are therefore best suited as transitional rather than permanent solutions. Bindel et al. (2017) show that backward compatibility in standards such as X.509, TLS, and S/MIME requires post-quantum alternatives to be embedded as non-critical certificate extensions.

Certificate lifecycles must be adapted by adjusting validity periods and issuance workflows to account for PQC key sizes and operational overhead. Short-lived certificates may reduce exposure but increase management complexity.

Computational overhead introduced by PQC can be mitigated through cryptographic accelerators, FPGAs, and other secure hardware platforms that support high-throughput signing and verification operations.

Standardization progress should be continuously monitored using resources from organizations such as NIST and ETSI. Participation in hackathons, interoperability testing events, and working groups supports early validation of deployment readiness.

Monitoring frameworks can be used to assess PQC performance and security in operational environments, while periodic audits verify compliance with operational and regulatory requirements. As monitoring tools are highly system-specific, their deployment must be tailored to the operational characteristics and risk profile of each PKI.

7. Conclusion

This paper examined the migration of time- and security-critical Public Key Infrastructures to post-quantum cryptography from a systems and PKI-engineering perspective. Using SWIM and C-ITS as representative case studies, we showed that PQC migration in such environments is constrained less by cryptographic correctness than by operational requirements related to latency, bandwidth, interoperability, and regulatory compliance. Our analysis indicates that direct replacement of classical algorithms with PQC counterparts is generally infeasible in critical PKIs without structural changes to certificate handling and communication protocols. In particular, increased certificate and signature sizes, combined with verification-heavy workloads, can significantly affect end-to-end performance even when individual cryptographic operations remain computationally efficient. These effects are amplified in environments with short communication windows and strict timing guarantees.

As a result, hybrid cryptographic approaches emerge as a pragmatic transitional solution, despite introducing additional complexity and new risk vectors. Hybridization should be understood as an intermediate step that enables crypto-agility while standards and implementations mature. Migration must be phased, context-aware, and aligned with sector-specific standardization.

The contribution of this work is a constraint-driven synthesis of existing benchmarks, standards, and protocol specifications, that highlights feasibility limits and design trade-offs for PQC adoption in critical PKIs. This paper provides concrete guidance for system designers and policymakers tasked with planning long-term cryptographic transitions in security-relevant infrastructures.

Future work should focus on system-level validation of PQC-enabled PKIs under realistic traffic conditions, as well as on the evolution of standards that better accommodate multi-algorithm operation and cryptographic agility. Such efforts will be essential to ensure that the transition to post-quantum cryptography strengthens, rather than compromises, the safety and reliability of critical digital infrastructures.

Acknowledgements

This work has been funded by the European Defence Fund research and innovation project SESIOP, under the programme grant agreement No. 101070052, and Business Finland project Beyond the Limits of Post-Quantum Cryptography, under the agreement No. 100/31/2024.

Ethical Clearance: Ethical clearance was not required for this research.

AI Declaration: Copilot AI has been used for language refinement.

References

Aardal, M.A., Adj, G., Aranha, D.F., Basso, A., Martínez, I.A.C., Chávez-Saab, J., Santos, M.C.-R., Dartois, P., Feo, L.D., Duparc, M., Eriksen, J.K., Fouotsa, T.B., Filho, D.L.G., Hess, B., Kohel, D., Leroux, A., Longa, P., Maino, L., Meyer, M., Nakagawa, K., Onuki, H., Panny, L., Patranabis, S., Petit, C., Pope, G., Reijnders, K., Robert, D., Henríquez, F.R., Schaeffler, S. and Wesolowski, B. (2025). SQISign algorithm specifications and supporting documentation version 2.0. https://sqisign.org/spec/sqisign-20250205.pdf.

Alagic, G., Apon, D., Cooper, D., Dang, Q., Dang, T., Kelsey, J., Lichtinger, J., Liu, Y.-K., Miller, C., Moody, D., Peralta, R., Perlner, R., Robinson, A. and Smith-Tone, D. (2022). Status report on the third round of the NIST post-quantum cryptography standardization process (NIST IR 8413-upd1). https://doi.org/10.6028/NIST.IR.8413-upd1.

Alagic, G., Bros, M., Ciadoux, P., Cooper, D., Dang, Q., Dang, T., Kelsey, J., Lichtinger, J., Liu, Y.K., Miller, C., Moody, D., Peralta, R., Perlner, R., Robinson, A., Silberg, H., Smith-Tone, D. and Waller, N. (2024). Status report on the first round of the additional digital signature schemes for the NIST postquantum cryptography standardization process (NIST IR 8528). https://doi.org/10.6028/NIST.IR.8528.

Aumasson, J.-P., Bernstein, D.J., Beullens, W., Dobraunig, C., Eichlseder, M., Fluhrer, S., Gazdag, S.-L., Hülsing, A., Kampanakis, P., Kölbl, S., Lange, T., Lauridsen, M.M., Mendel, F., Niederhagen, R., Rechberger, C., Rijneveld, J., Schwabe, P. and Westerbaan, B. (2022). SPHINCS+ submission to the NIST post-quantum project, v.3.1. https://sphincs.org/data/sphincs+-r3.1-specification.pdf.

Barnes, R., Hoffman-Andrews, J., McCarney, D. and Kasten, J. (2019). Automatic Certificate Management Environment (ACME). RFC 8555. https://www.rfc-editor.org/info/rfc8555.

Beullens, W., Campos, F., Celi, S., Hess, B. and Kannwischer, M.J. (2024). MAYO round 2 version. https://pqmayo.org/assets/specs/mayo-round2.pdf.

Bindel, N., Herath, M., McKague, M. and Stebila, D. (2017). Transitioning to a quantum-resistant public key infrastructure. In *PQCrypto 2017*, Utrecht, The Netherlands, 26–28 June. Springer, pp. 384–405.

Bos, J.W., Bronchain, O., Ducas, L., Fehr, S., Huang, Y.-H., Pornin, T., Postlethwaite, E.W., Prest, T., Pulles, L.N. and van Woerden, W. (2025). HAWK version 1.1 (February 5, 2025). https://csrc.nist.gov/csrc/media/Projects/pqc-dig-sig/documents/round-2/spec-files/hawk-spec-round2-web.pdf.

ETSI (2019a). *ETSI EN 302 637-2 V1.4.1: Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service.*

ETSI (2019b). *ETSI EN 302 637-3 V1.3.1: Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralized Environmental Notification Basic Service.*

ETSI (2021). *ETSI TS 103 097 V2.1.1: Intelligent Transport Systems (ITS); Security; Security Header and Certificate Formats; Release 2.*

ETSI (2022). *ETSI TS 102 941 V2.2.1: Intelligent Transport Systems (ITS); Security; Trust and Privacy Management; Release 2.*

ETSI (2023). *ETSI TR 103 949 V1.1.1: Quantum-Safe Cryptography (QSC) Migration; ITS and C-ITS Migration Study.*

European Commission (2021). *Commission Implementing Regulation (EU) 2021/116 of 1 February 2021 on the establishment of the Common Project One supporting the implementation of the European Air Traffic Management Master Plan.* Official Journal of the European Union.

European Commission (2022). *Commission Decision of 12.8.2022 approving the SESAR deployment programme 2022.* https://www.eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022D1344.

EUROCONTROL (2022a). *SWIM Common PKI: Common PKI Specifications (D2.1).* Brussels: EUROCONTROL/SESAR Deployment Manager.

EUROCONTROL (2022b). *SWIM Common PKI and Policies & Procedures for Establishing a Trust Framework: Certificate Policy, Annex A3.a of D1.2 Final Trust Framework.* Brussels: EUROCONTROL/SESAR Deployment Manager.

ICAO (2018). *Manual on System-Wide Information Management (SWIM).* Montreal: ICAO (Doc 10039).

ITU (2025). Recommendation X.509: The directory – authentication framework. https://www.itu.int/rec/T-REC-X.509.

Kampanakis, P., Panburana, P., Daw, E. and Van Geest, D. (2018). The viability of post-quantum X.509 certificates. *Cryptology ePrint Archive*, Paper 2018/063.

Koblitz, N. (1987). Elliptic curve cryptosystems. *Mathematics of Computation*, 48(177), pp. 203–209.

Nikula, S., Halunen, K. and Vallivaara, V. (2022). Quantum-safe signing of notification messages in intelligent transport systems. *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering*, 448, pp. 11–25.

NIST (2024a). *Module-Lattice-Based Digital Signature Standard (FIPS 204).* https://doi.org/10.6028/NIST.FIPS.204.

NIST (2024b). *Stateless Hash-Based Digital Signature Standard (FIPS 205).* https://doi.org/10.6028/NIST.FIPS.205.

NIST (2024c). *Module-Lattice-Based Key-Encapsulation Mechanism Standard (FIPS 203).* https://doi.org/10.6028/NIST.FIPS.203.

NIST (2024d). *NIST IR 8547: Transition to Post-Quantum Cryptography Standards, Initial Public Draft.* https://nvlpubs.nist.gov/nistpubs/ir/2024/NIST.IR.8547.ipd.pdf.

Ounsworth, M., Gray, J., Pala, M. and Klaußner, J. (2024). Composite ML-DSA for use in Internet PKI. https://www.ietf.org/archive/id/draft-ounsworth-pq-composite-sigs-13.html.

Sarıbas, S. and Tonyalı, S. (2022). Performance evaluation of TLS 1.3 handshake on resource-constrained devices using NIST's third-round post-quantum key encapsulation mechanisms and digital signatures. In *2022 7th International Conference on Computer Science and Engineering (UBMK)*, pp. 294–299.

SESAR-NextGen (2018). *NextGen – SESAR State of Harmonisation: Third Edition.* https://sesar.eu/sites/default/files/documents/StateOfHarmonization_Ed3.pdf.

Yoshizawa, T. and Preneel, B. (2023). Post-quantum impacts on V2X certificates – already at the end of the road. In *2023 IEEE 97th Vehicular Technology Conference*.

From Water Plants to Nuclear Reactors: Mapping SCADA Vulnerabilities in Small Modular Energy Systems

Shreyas Kumar, Aleksander Alvarez, Jashanjodh Bajwa and Shrutwik Muppa
Texas A&M University, College Station, USA

shreyas.kumar@tamu.edu
alekalvarez@tamu.edu
jashanjodhb@tamu.edu
shrutwik@tamu.edu

Abstract: Supervisory Control and Data Acquisition (SCADA) systems are responsible for helping to manage a large portion of the industrial process, which keeps the economy running. In Small Modular Reactors (SMRs), SCADA helps with monitoring reactor conditions but is responsible for the careful variable control, which keeps reactors running optimally. Due to the far-reaching benefits of SCADA involvement, when there is a SCADA related attack, the consequences are also far-reaching. SMRs rely even more heavily on interconnected digital systems, meaning similar attacks could destabilize reactors, shut down industrial sites, or disrupt critical facilities such as refineries and military bases. This paper takes a closer look at SCADA vulnerabilities in the context of SMRs. Through analysis of the National Vulnerability Database, we compiled likely attacks on SMRs and evaluated them for threat severity and likelihood. Using trend analysis, topic modelling, and economic impact assessment, we show how these weaknesses could affect energy infrastructure and, more importantly, the many industries that rely upon it. Our findings underline the urgent need to strengthen the infrastructure defences, specifically in the realm of small modular nuclear reactors.

Keywords: Supervisory control and data acquisition, Industrial sabotage, Infrastructure protection, Modular nuclear reactors, Topic modelling

1. Introduction

Supervisory Control and Data Acquisition (SCADA) software has become increasingly implemented in industrial processes over time and, in the context of this study, is an integral part of the operation of Small Modular Nuclear Reactors (SMRs). SCADA software functions as the intermediary between human controllers and automated machinery and provides a platform to monitor and adjust variables in real-time. However, due to this integral role, vulnerabilities in SCADA software have great consequences for critical and sensitive infrastructure. A notable example of the potential far-reaching effects is a 2021 cyber-attack that occurred in Oldsmar, Florida. In the attack, the attacker attempted to exploit a simple remote access vulnerability to access a SCADA operator terminal in a water treatment facility, which could potentially have led to thousands of deaths (Davis, 2024). In nuclear reactors, SCADA is used to keep track of numerous variables that influence the fission and electricity-generating process and additionally coordinating between SMRs. SCADA monitors and adjusts these variables to keep the reactor stable and functioning optimally (Wang, 2024). To monitor and coordinate SMRs, increased numbers of digital devices will be needed, and these open pathways for attack (IAEA, 2023). With the growing popularity of SMRs, the cybersecurity issues become more urgent. This paper analyses SCADA vulnerabilities in the context of SMRs and explores their economic ramifications.

2. Related Work

Kaura, Sindhwani, and Chaudhary analysed the impact of cyber threats on SCADA systems and Industrial Control Systems. Focused on major real-world incidents such as Maroochy Water, Dragonfly, and TRITON. They introduced a three-tier system model: CAT-I, which involves direct damage in operations; CAT-II, attacks that cause potential disruption; and CAT-III, which focuses on espionage and low-level threats. The authors point out trends that include the rising role of state actors and the persistent APT campaigns. They emphasize the development of malware designed specifically for industrial control. The authors also mention phishing and social engineering as vulnerabilities on such systems. This is significant within the realm of SMRs due to the potential that the energy industry poses as an attractive target in widespread economic sabotage. In many of the uses of SMRs explored later such an attack would lead to great impact.

Daryabar et al. (2012) presented an extensive analysis of SCADA security from multiple threat sources such as insider actors, terrorists, malware developers, hackers, and hobbyists. SCADA systems are exposed as they rely on real-time functionality and wide-area connectivity, making them vulnerable to delays, disruptions, and unauthorized remote access that can directly impact critical operations. The authors stress the need for a solution for prolonged downtime as systems do not run due to their real-time nature. To mitigate such

vulnerabilities, the authors provide a couple of tests and advancements, such as red-team exercises, penetration testing, honeynet deployments, and stronger monitoring systems. Combining the measures provided by the authors would enable a protected and structured model.

3. Methodology

This study confronts the issue of SCADA vulnerabilities in relation to SMRs in a novel combination of various approaches.

3.1 Filtering

To quantify trend in vulnerability analysis, we rely on the National Vulnerability Database, which is an enriched list of CVE datapoints with CVE data combined with CVSS, CWE among other enriching data (NIST, n.d.). Common Vulnerability Scoring System CVSS is a standardized metric that provides a way to qualitatively understand a vulnerability's impact (NIST, 2022). Common Weakness Enumeration CWE is a collection of weakness categories that allows you to classify vulnerabilities by type (CWE, 2023).

To mine the data from the national vulnerability database, zip files containing the corresponding JSON files for 2014-2024 were used (nvd.nist.gov, n.d.). CVSS versions 2 and 3 was used. Through scripting, we identified data points related to SCADA through filtering vulnerability descriptions for keywords related to SCADA technologies. CWE, CPE, and CVSS information was then extracted from these entries.

3.2 Topic Modelling

To automate trend analysis in a way to find patterns hidden via the generalization of the CWE's vulnerability classification, topic modelling was applied to find common themes throughout Vulnerability Summary Text. First, this involved preprocessing steps that removed punctuation, stopping words and other non-descriptive common words. Then, this processed text data was plugged into a Latent Dirichlet Allocation (LDA) model. To maximize the quality of the LDA analysis for every year, a human supervisor finetuned parameters such as topic count and word count, leading to the results featured in our table.

3.3 Economic Analysis

To create an outlook for current market growth, the World Nuclear Association's database was analyzed for the status of all current SMRS. These results were filtered to only include the currently operational SMRs.

To examine the market outlook, the analysis was focused on the current quantity and quality of investments into SMR projects that had passed into the construction phase.

To create a model for the economic loss from nuclear reactors, data was acquired from the U.S. Energy Information Administration on the wholesale price of electricity. The number of reactors and total output in MWh in the U.S. was gathered from the World Nuclear Association to create an average output and gross profit figure per MWh.

The costs were incorporated by considering the three largest expenditures of fuel, capital, and operating costs. These costs were prioritized because they represent the bulk of the costs in SMRs. Smaller, miscellaneous expenses such as regulatory fees were not included in the costs, leading to an underestimation.

4. Results

4.1 SCADA Vulnerability Quantitative Trends

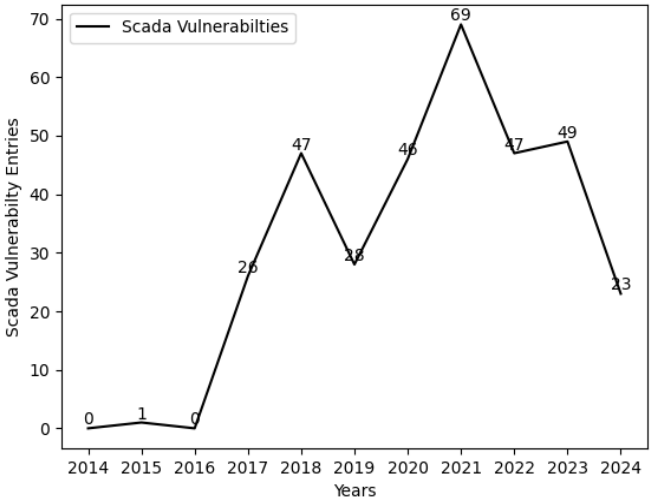


Figure 1: Increasing Trend in SCADA Vulnerabilities on the NVD

Figure 1. was created by calculating the number of SCADA vulnerabilities isolated per year. As seen from the graph, there is a general upward trend in vulnerabilities being identified as SCADA-based or SCADA adjacent vulnerabilities over the time of 2014 to 2024. There is a downward trend in vulnerabilities three times throughout the ten-year time but when comparing the start and end of the ten-year period, there is a definite trend upwards. Although compared to the total number of reported vulnerabilities in the national vulnerability database the SCADA-related vulnerabilities are quite small, the identified vulnerabilities are not trivial. SCADA vendors are relatively low in number, meaning that there are relatively few products that dominate the market meaning that the low number of vulnerabilities has a high impact. With the diversity featured in the different vulnerabilities, there is a significant risk to public infrastructure as well as energy generation.

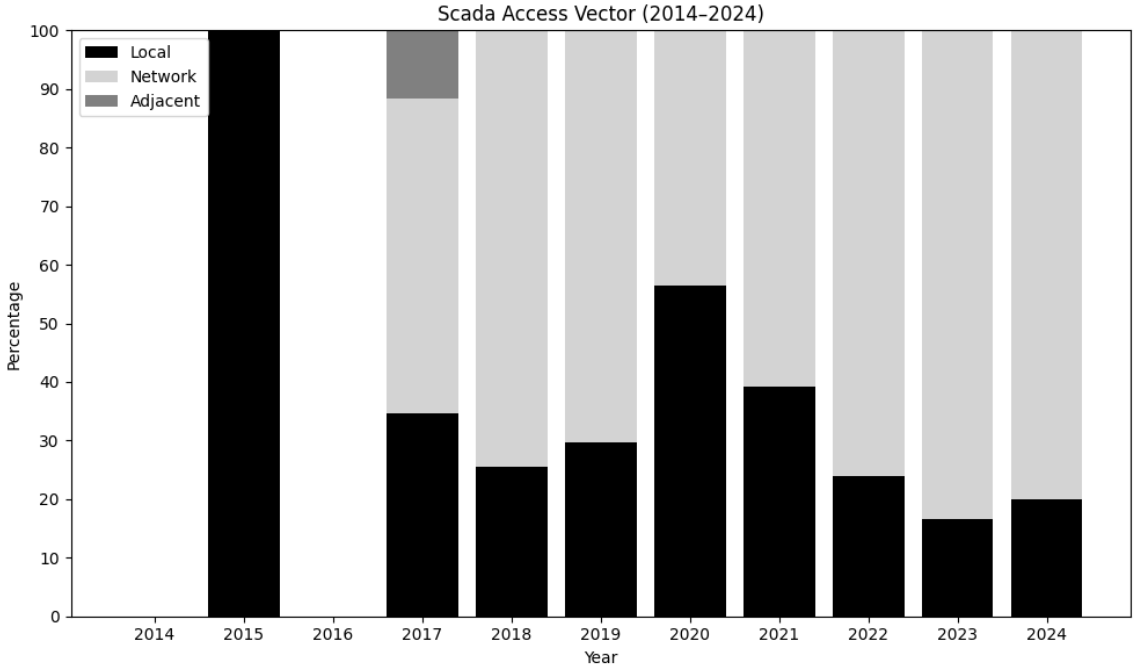


Figure 2: Attack Vectors of SCADA Vulnerabilities (2014-2024)

It is important to note from **Figure 1.** that there are several years where there is a low sample size, which is reflected in all trends drawn later. As seen from **Figure 2.**, the adjacent vector is the least prevalent vector type. This includes Bluetooth and attacks over the same network. What is most concerning with the case of modular nuclear reactors is network-based accounts, which happen to be the most common type of attack vector. Due to the nature of the use of modular nuclear reactors, most of them happen to be in very controlled environments where local attacks would only be possible if done by a previously authorized employee. While this certainly is a possibility that should be considered, the threat of a malicious remote hacker is one that should be the greater threat, as such an attacker would have anonymity and could reasonably cause havoc and potentially profit without consequence. Therefore, it makes it very concerning that these are the most common access vectors, which are remote. Combined with the prevalent categories for the Common Weakness Enumeration data, realistic possible attacks on modular nuclear reactors can be thought of and then planned. For example, with a hard-coded credentials vulnerability, a remote attacker could possibly sign in as an administrator remotely, thus potentially triggering safety measures that could affect reactor power output.

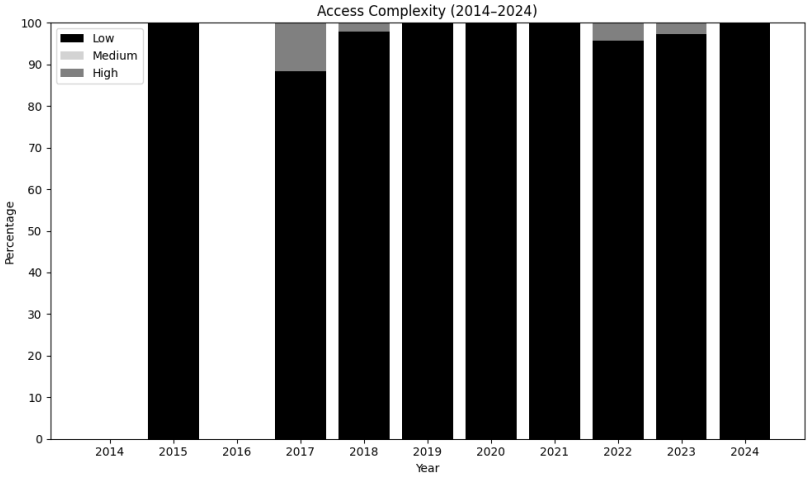


Figure 3: Access Complexity of SCADA vulnerabilities 2014-2024

From **Figure 3** all vulnerabilities are classified in three access complexity categories. Low complexity attacks pose a great security risk due to how they tend to be much easier to replicate, and these are unfortunately most common. For these reasons, they tend to be the vulnerabilities that have the maximum benefit in studying.

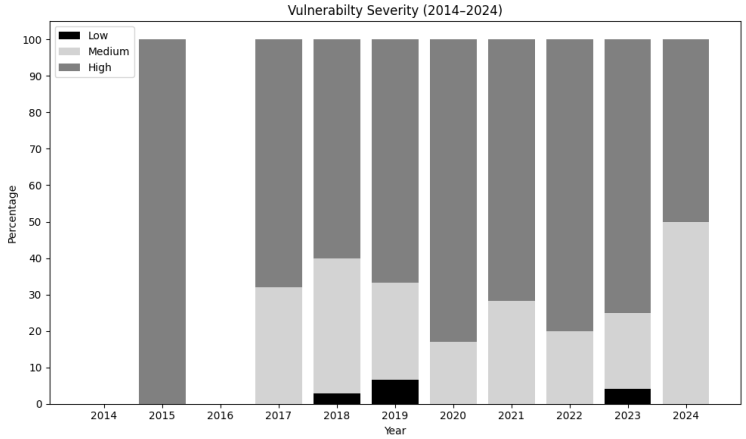


Figure 4: Vulnerability Severity 2014-2024

The National Vulnerability Database for its entries defines three scores to describe the severity risk, which is a standardized measure of the impact a vulnerability has on the user which interacts with it. In **Figure 4.** most vulnerabilities that are SCADA-related fall into the category of high severity. This is a concerning trend, as this metric is a measure of how the vulnerability affects the user. Since the user in this case is SCADA users, this means that most vulnerabilities in this field have a relatively high effect on these industrial systems. With the

application of SCADA software for remote monitoring and controls for modular nuclear reactors, the consequences of this trend are clear and quite important.

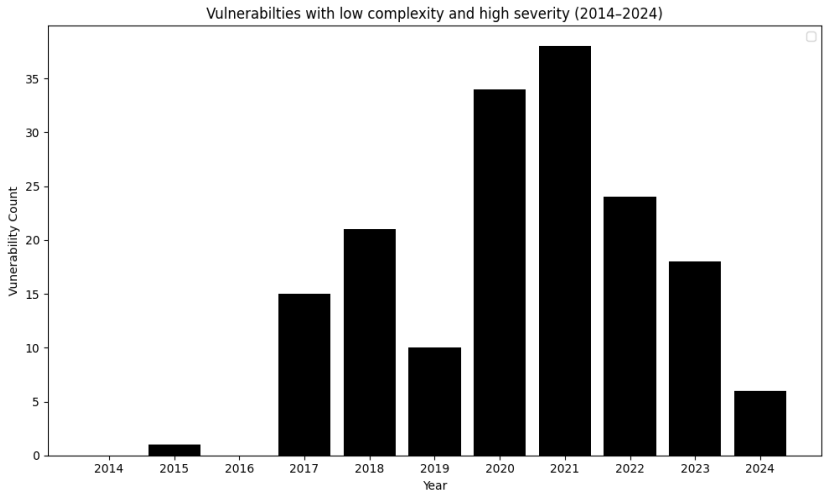


Figure 5: Trends of most serious vulnerability types 2014-2024

Vulnerabilities that are of prime importance are those whose complexity is low, along with a high severity.

From **Figures 1** and **5**, the trend with low complexity and high severity graph aligns closely with the SCADA in terms of the time figure. From this, the conclusion can be drawn that there tends to be around a certain percentage of vulnerabilities that are higher in severity and lower in complexity. This is significant because, as noted before, the low complexity vulnerabilities encourage a higher number of attacks before the solution to a vulnerability is found, and high severities increases the possibilities of consequences that malicious actors can implement on their targets. These vulnerabilities are the key ones that security professionals should analyze to create procedures and physical redundancies that can serve to immunize future reactors.

4.2 Categorical SCADA Vulnerability Analysis

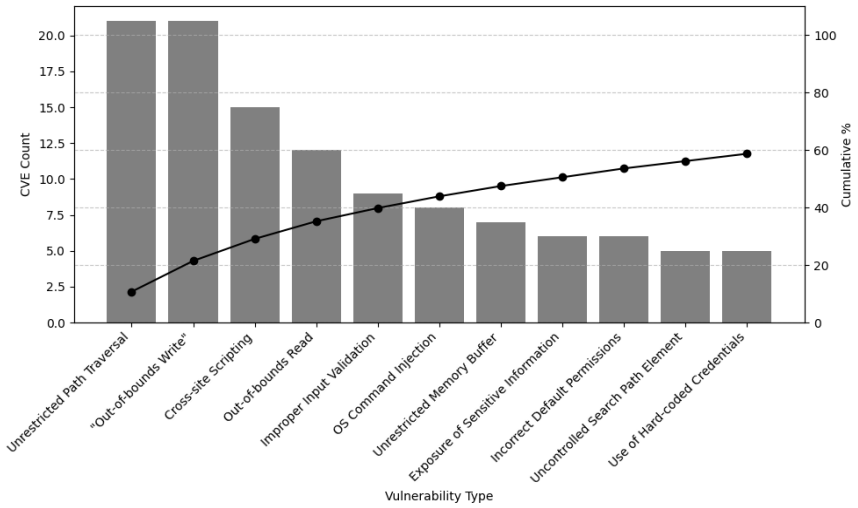


Figure 6: CWE Classification

Using the Common Weakness Enumeration data recorded in the NVD vulnerabilities, which were identified as SCADA-related vulnerabilities, the vulnerabilities were sorted by Category type. Using this sorted data, the Prevalent Categories graph was created. It is important to note that the noted category names are shortened forms of the CWE names. As seen from the graph, there is no one category with a clear domination of the dataset, yet there are some that are more prevalent than others. For example, the first eleven vulnerability categories, which are the ones featured in the graph, are the categories into which sixty percent of the recorded SCADA vulnerabilities fall into. The prevalence of some of these categories has obvious ramifications for Nuclear Reactors, as unrestricted path traversal regarding a SCADA system could lead to an attacker having

access to nearly all aspects of a remote system. Out of bounds write and out of bounds read both have similar implications, with out of bounds writes potentially leading to registers used by controlled devices being written two thus impacting the integrity of a SCADA system, for example parameters concerning reactor temperature could be changed, thus allowing an attacker to remotely cease a reactor's power production. For out-of-bounds read errors, this can affect the confidentiality of the information of the reactor. Such an attack could be used to steal admin credentials to enable a later system-critical attack on a SCADA energy system. One category with the lowest relative prevalence displayed on the graph is hard-coded credentialing, which, for obvious reasons, could lead to drastic consequences for modular nuclear reactors that utilize SCADA systems.

Table 1: Common Themes From Topic Modelling 2017-2020

2017	2018	2019	2020
Spider Control	Web Access Scada	Buffer Overflow	Local Execute
Directory Traversal	Remote Code Execution	Authentication	Directory Upgrade
Schneider Solutions	File Exfiltration	arbitrary reference	Installation Privilege
Uncontrolled Search	Disclosure	Denial of Service	IGSS
Cross Site Scripting	Authentication	Information Read	Remote Data Connect
Citect	Remote Port		Packet
Overflow Issue	Control Access		Server
Resource Consumption	Escalated Privilege's		Code Execution
	SQL Injection		User Authorization
	PLC Issue		Out of Bounds Read

Table 2: Common Themes From Topic Modelling 2021-2024

2021	2022	2023	2024
Signature Forge	User Privileges	Hardcoded Error	File Access
Endpoint Access	Buffer Overflow	SCADA Gateway	Firmware
Code Execution	User Authentication	File Overwrite	Code Injection
User Bypass	Input Validation	Code Execution	Logging
Passwords	Denial Of Service	Shell	User Authentication
Server Request		Certificate	
		Path Traversal	
		Plaintext Password Storage	

The value of the LDA Analysis can be seen in how quite a few of the captured themes overlap with the common CWE categories. This suggests that despite the limitations of the LDA analysis on smaller documents it still provides a reliable automated analysis. Additionally, there a more specific themes which would not have been captured by the CWE analysis such as specific attack vectors which shall prove to be valuable.

4.3 SMR Market Analysis

The current market of SMRs is quite small due to the relative recency of the technology, with only two operating reactors in China and Russia (World Nuclear Association, 2025a). This is changing rapidly, as billions are invested. The Korean SMR SMART 100 has been approved by Korea's Nuclear Safety and Security Commission in September 2024, after nearly 4 years of waiting (World Nuclear News, 2024). With the cost and risk associated with nuclear power in general, investments come in bursts, as was the case with the Darlington New Nuclear Project, which invested 15.1 billion USD into four SMRs at once. The reactors are to be built in phases, with the final reactor finishing in the mid-2030s. The cost of these reactors is also projected to be progressively cheaper as the new reactors are built, with the final being around 1/3 cheaper than the first (World Nuclear News, 2025). In comparison, the cost of Vogtle Units 3 and 4 in Georgia, United States, which are the latest reactors built in the United States, is estimated to be over 30 billion USD (U.S. Energy Information Administration, 2024). When comparing output, the investment is quite similar between the two

but the scalability for SMRs can be superior due to lower upfront capital costs, leading to a market advantage. The market of SMRs can be defined by the categories of geographical advantage, industry application, infrastructure requirements, and, most importantly, scalability (Gaster, 2025). The niche group of geography, industry, and infrastructure is the smaller share of the market where SMRs have the advantage of being able to supply remote areas, industrial needs for heat, or areas with poor infrastructure, instead of other energy options. The larger group is the scalability factor of SMRs for general electricity production (Lokhov, Cameron, Sozonuik, 2013). The advantage of being able to mass-produce reactors and interconnect them to create an easily scalable power plant is a huge advantage for SMRs. Thus, the market for SMRs is dependent on the extent to which this property is employed. While scalability has shown to decrease cost, the hindrances from time costs and initial investments may lead to slow implementation. There are currently over 80 SMRs in some phase of development, and the market outlook for SMRs is dependent on how many of these projects make it to fruition (International Atomic Energy Agency, 2025). SMR technology has immense potential, but its growth will only be realized in the coming decades.

4.4 Economic Loss Model

To estimate losses, we will look at the profits of the Average nuclear reactor and analyze the per-GWh profit. These numbers are based on the current U.S. market and nuclear capacity. There are currently 94 operable reactors with a total of 781,945 GWh of electricity generation in 2024 (World Nuclear Association, 2025b). Based on this, there is an average of 8318.56 GWh of electricity produced on average by each reactor per year. 22.79 GWh is produced every day, or 949.61 Megawatts per Hour. During 2024, the average sale price of wholesale electricity was $44.82 per megawatt-hour of electricity (U.S. Energy Information Administration, 2025). The prices vary by the energy market and when in the year they are sold, but we can take the average of all markets over all time, as we did with the nuclear capacity. With these numbers, we can come up with a gross income for the average reactor and by megawatt, but for profitability, we need expenses. There are several factors that can push profitability down, such as licensing fees, waste management, and other smaller costs. To simplify, we will focus on the big expense that includes fuel, capital costs, and operating costs. In the year 2023, the average cost in the U.S. for fuel was $5.32, for capital was $7.06, and for operating was $19.38 per MWh (Nuclear Energy Institute,2025). This gives us a total of $31.76 per MWh in costs. Now, to create a profit per MWh, we take out $31.76 of the average cost of electricity at $44.82 and get $13.06 in profit per MWh produced. To apply this to the average nuclear reactor in the United States currently, it would mean a profit of about $12,401.91 per hour or about $297,645.76 per day. These numbers are based on averages of both revenues and costs, so there will be an error in the calculations, and profit will be significantly higher for reactors who have already paid off their construction costs. Keeping this in mind, the profit will be significantly higher for some reactors.

4.5 Projected Losses for SMRs

As we have defined in earlier sections, one of the most important attributes of SMRs is their scalability. Though SMRs are typically rated up to 300MW, like the Korean SMART 100, the scalability allows for the potential for vast networks of SMRs working together. In the case of Canada's Darlington New Nuclear Project, 4 SMRs together would add up to 1200 MW of energy. This scalability also runs both ways when looking at the vulnerability of these reactors to cyber-attacks, especially SCADA. With the average nuclear reactor profiting nearly $300,000 a day, reactors being taken down are costly. This is only exacerbated by the scale at which SMRs might be deployed. If SMRs can reach the level of production that makes them competitive with other forms of power, security will need to be taken into even higher consideration. Though currently nuclear plants already have tight security procedures, the quantities involved in scaling SMR technology may require security to become even more rigorous. Thus, cyberattacks will have dire consequences as SMR scalability leads to plants having even higher outputs than the average reactor currently, leading to losses of magnitudes far greater than the $300,000 per day we might see today.

4.6 Scenario-Based Losses

Earlier sections provided a general estimation of losses from reactor downtimes in monetary terms. This section aims to look at specific scenarios and their impact and probability of occurrence but in relative terms. We will investigate how different vulnerabilities translate to physical action and economic impact relative to each other. We will not investigate exact probabilities, which can be difficult to quantify, and instead focus on relative estimations of common vulnerability types. We will choose one representative vulnerability from three tiers of likelihood. From **Figure 6**, path traversals are high likelihood, OS Command injections are medium likelihood, and hard-coded credential vulnerabilities are low likelihood. Economic impact can be highly varied,

even in specific cases. Thus, we will look at the broad impact and label tiers to the economic impact. Path traversal attacks aim to access files that are outside of the web root and can have varied consequences. The most extreme of these is the crash of the program, where the attacker can edit or delete critical files (Common weakness enumeration, 2024b). Depending on the criticality of the impacted system and the protections implemented, this vulnerability can range from no impact to shut down for long periods of time. This leads to a moderate economic impact, as the severity of vulnerability is not always critical. OS command injections, on the other hand, function through the lack of input validation. These attacks include the attackers adding their own code to be executed by the system (Zhong, n.d.). Like path traversal attacks, these vulnerabilities have a varied impact and share the same extreme of allowing for the deletion of files. Due to identical reasons, this vulnerability has the potential for moderate economic impact. Lastly, hard-coded credential attacks involve hard-coded passwords being used to gain access to a system (Common weakness enumeration, 2024a). This essentially leads to an attacker having privileged access to the complete system, depending on the credentials. Due to the span of access, this is the most critical vulnerability out of the three discussed. Correspondingly, this vulnerability can lead to the highest economic impact, as an attacker can create damage on a higher level of magnitude. Overall, these specific vulnerabilities vary in their probability of occurrence, but their economic impact is typically moderate to high.

5. Discussion

While the numerous aspects of this analysis have been done before, novel and actionable insights from this paper when a holistic view is taken. Through a combination of the NVD quantitative and qualitative analysis in addition to the economic analysis, a complete picture of the landscape to be combatted can be formed.

5.1 Overall Trends Analysis

The National Vulnerability Database Analysis period of 2014 to 2024 displays a concerning trend about SCADA vulnerabilities. From **Figure 1**. SCADA-related vulnerabilities have generally increased over this period. Although they comprise a small fraction of the total vulnerability count with how a large market share of SCADA services originate from very few companies (NERC), this trend should be cause for alarm. If most SCADA software's originate from relatively few companies, then a relatively small number of vulnerabilities, such as the ones studied in this paper, with a diverse range of origins, which can be seen in the vendors that pop up in the thematic analysis, will have wide reaching and concerning affects.

From **Table 1** as well as the CWE Categorical Classification, most vulnerabilities fall into the categories of Privilege Escalations and Code Execution. This second category is of great interest due to the high potential impact it has. Assuming all safety measures occur as planned, the worst-case scenario for remote code execution would be the temporary shutdown of the Reactor. For example, CVE-2019-10980 is one of the high severity low complexity vulnerabilities allowing project files to remotely execute code. Disregarding the secondary effects which a reactor shutdown would have, this could cost around $300,000 if turnaround time was less than a day, which is an unrealistic time frame. It would be advisable for nuclear SCADA software to separate data manipulation scripting from sensor reading to prevent customization on the part of the user as much as possible. This would decrease the risk of such a remote attack.

Through a combination of economic analysis, quantitative analysis, and categorical analysis, further headways in securing the future of Modular Nuclear Reactors can be made.

5.2 Privilege Escalations

From both the categorical analysis and the topic modelling, one persistent and common threat to SMRs is the category of privilege escalation. Beyond interfering with the operation of a modular nuclear reactor, this also poses a risk to the confidentiality of the plant's application. It is important to remember that many SMRs are intended for usage with critical industrial sites and military sites. In these applications, information regarding power usage has obvious value to cyber adversaries and thus is an additional impact that must be studied/considered.

Beyond information disclosure concerns, privilege escalation has measurable concerns when it comes to affecting the operational integrity of the SMR. The CWE analysis demonstrates quite a few categories relating to privilege escalation are of particular concern, and looking at the topic model, more specific themes to SCADA include SCADA gateways. When analysing vulnerabilities mined over the relevant time frame throughout the year 2023, a pattern of insecure privilege practices, which could pose great challenges to the operations of the SMR.

For example, CVE-2023-39460 would allow an attacker to potentially create an unauthorized file in the SCADA system. If, hypothetically, an attacker were to exploit this to create an executable capable of affecting reactor function this would have a greater than $300,000 impact.

From this pattern of vulnerabilities, it is advisable that for applications involving SMRs, practices of least privilege and non-customization should be adopted, thus strengthening the security of the system. Additionally, physical input should be required for actions that would result in a shutdown.

5.3 Vulnerabilities of Key Interest

Throughout this paper, vulnerabilities of note have been collected for their identifying characteristics of being both High in severity and low in complexity. It is this set that, when studied, it reveals pertinent answers regarding strengthening Modular Nuclear Reactors. One vulnerability to highlight is CVE-2024-10313. This vulnerability in SpiderControl SCADA allowed a malicious energy management system project template file to write files to arbitrary locations, which could lead to consequences such as overwriting files, system paralysis, or adversaries gaining remote control of equipment (NIST, 2024). With consequences of this severity, it is critical for fail-safes to be implemented and required across all reactors. A digital solution could include a virtual testing environment that analyses the file for arbitrary writes before clearing it. This is, however, not a full solution and would need to be used in conjunction with other safety protocols. Using a virtual machine (VM) can isolate the overwrites, but attackers may still cause damage if the VM is not used in a segmentation firewall, which isolates the VM from the rest of the network. These are some general measures that can be used to limit the impact of these vulnerabilities. While they are good practices, it is still crucial to maintain the latest cybersecurity measures, as attacks are constantly evolving.

5.4 Limitations

There are some important limitations that are important to note from this papers analysis.

As stated, before trends derived from the NVD are oftentimes and underestimate of the true landscape. This is because the National Vulnerability Database is dependent on voluntary reporting from researchers and developers. It is important to utilize the NVD as it remains the most reliable way to compare vulnerabilities across different platforms. Additionally, LDA analysis is known for having noisy results when done over small text datasets, but as seen in the similarities between some topics covered in the categorical analysis and the thematic analysis as well as how more pertinent trends to SCADA can be found using LDA it has its place in advancing this analysis.

As for the limitations of the economic analysis, it was mentioned before that several types of expenses for nuclear reactors are not included in the costs. This is due to the amount of variation and the weight of the miscellaneous expenses, such as fees. These expenses can vary between states and countries, which makes it difficult to account for all of them. The fact that these expenses are, comparatively, smaller than the three categories that are accounted for also reinforced the decision to omit these expenses. This does, however, make the costs an underestimate and the profits an overestimate. It is also important to note that in the process of creating an average of the wholesale price for electricity, this study simplifies the prices in several markets in the United States into a single average. This takes away some variation, and thus the study may not align completely with prices in different markets.

6. Conclusion

This paper presents a quantitative and qualitative assessment of the risk environment in which Modular Nuclear Reactors find themselves, in addition to an economic analysis to quantify these results. The quantitative analysis depicts an increasingly hostile environment where attackers can increasingly utilize network-based attacks to cause cyber terror at a distance. The qualitative analysis reveals a thematic emphasis on privilege escalation and its subsets. This has a specificity that would not be available relying solely on CWE categories. This, combined with the direct economic impact this has been shown to cause, is why action in this area is required.

6.1 Key Takeaways

For the developers of SCADA systems, which are used in Modular Nuclear Reactors, greater strengthening is needed around Permissions. These includes greater inclusions of Sandbox architecture as well as isolation of client-side scripting. For the integrators of SCADA with Modular Nuclear Reactors care should be taken to have physical redundancy for critical operations and sequences. If there is physical verification of certain actions

shut as reactor shutdown, then attacks would have to increase in sophistication. For research professionals' interest should be applied to computing the total indirect cost of this category of attacks, as well as vulnerability evolution.

Ethics declaration: As seen in Section 2 there were no human subjects, and all analysis was done on public data therefore no ethical board was required.

AI declaration: AI was not used throughout this paper besides the LDA analysis which we conducted.

References

CWE (2023). *CWE - New to CWE*. [online] cwe.mitre.org. Available at: https://cwe.mitre.org/about/new_to_cwe.html.

CWE (2024a). Common weakness enumeration. [online] cwe.mitre.org. Available at: https://cwe.mitre.org/data/definitions/798.html

CWE (2024b). Common weakness enumeration. [online] cwe.mitre.org. Available at: https://cwe.mitre.org/data/definitions/22.html

Daryabar, F., Dehghantanha, A., Udzir, N.I., Sani, N.F.M. and Shamsuddin, S.B. (2012) "Towards secure model for SCADA systems", *2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec)*, pp. 312–317. [online] IEEE. Available at: https://doi.org/10.1109/CyberSec.2012.6246111 [Accessed 30 Sep. 2025].

Davis, R. and Keskin, O.F. (2024). "Cyber Threat Modeling for Water and Wastewater Systems: Contextualizing STRIDE and DREAD with the Current Cyber Threat Landscape." doi:https://doi.org/10.1109/sieds61124.2024.10534706.

Gaster, R. (2025) "Small Modular Reactors: A Realist Approach to the Future of Nuclear Power", [online], Itif.org, Available at: https://itif.org/publications/2025/04/14/small-modular-reactors-a-realist-approach-to-the-future-of-nuclear-power/.

Iaea.org. (2023). *Securing Digital Technologies of the Next Generation of Nuclear Reactors | IAEA*. [online] Available at: https://www.iaea.org/bulletin/securing-digital-technologies-of-the-next-generation-of-nuclear-reactors [Accessed 25 Aug. 2024].

International Atomic Energy Agency (2016) *Small Modular Reactors (SMR)*, [online], Iaea.org, Available at: https://www.iaea.org/topics/small-modular-reactors.

Kaura, C., Sindhwani, N. and Chaudhary, A. (2022) "Analysing the Impact of Cyber-Threat to ICS and SCADA Systems", *2022 International Mobile and Embedded Technology Conference (MECON)*, pp. 115–120. [online] IEEE. Available at: https://doi.org/10.1109/MECON53876.2022.9752425 [Accessed 30 Sep. 2025].

Lokhov, A., Cameron, R. and Sozonuik, V. (2013) *OECD/NEA Study on the Economics and Market of Small Reactors*, [online], Nuclear Engineering and Technology, 45, pp.701–706, doi: https://doi.org/10.5516/NET.02.2013.517.

NERC | Report Title | Report Date I Cyber Security Supply Chain Risks Staff Report and Recommended Actions. (2019). Available at: https://www.nerc.com/pa/comp/SupplyChainRiskMitigationProgramDL/NERC%20Supply%20Chain%20Final%20Report%20%2820190517%29.pdf [Accessed 26 Sep. 2025].

NIST (n.d.). *NVD - General*. [online] nvd.nist.gov. Available at: https://nvd.nist.gov/general.

NIST (2022). *NVD - Vulnerability Metrics*. [online] Nist.gov. Available at: https://nvd.nist.gov/vuln-metrics/cvss.

nvd.nist.gov. (n.d.). *NVD - Data Feeds*. [online] Available at: https://nvd.nist.gov/vuln/data-feeds.

Nuclear Energy Institute (2025) *Nuclear Costs in Context*, [online], Available at: https://www.nei.org/CorporateSite/media/filefolder/resources/reports-and-briefs/2024-Costs-in-Context_final.pdf.

U.S. Energy Information Administration (2024) *Plant Vogtle Unit 4 begins commercial operation*, [online], Available at: https://www.eia.gov/todayinenergy/detail.php?id=61963.

U.S. Energy Information Administration (2025) *Wholesale Electricity and Natural Gas Markets Data*, [online], Available at: https://www.eia.gov/electricity/wholesale.

Wang, Y., Chen, W., Zhang, L., Zhao, X., Gao, Y. and Venkata Dinavahi (2024). *Small Modular Reactors: An Overview of Modeling, Control, Simulation, and Applications*. *IEEE Access*, pp.1–1. doi:https://doi.org/10.1109/access.2024.3351220.

World Nuclear Association (2025a) *Small Modular Reactor (SMR) Global Tracker*, [online], Available at: https://world-nuclear.org/information-library/current-and-future-generation/small-modular-reactor-smr-global-tracker.

World Nuclear News (2024) *South Korean SMR design approved by regulator*, [online], Available at: https://www.world-nuclear-news.org/articles/south-korean-smr-design-approved-by-regulator.

World Nuclear News (2025) *Canada's first SMR project: How is CAD20.9 billion cost calculated?,* [online], Available at :https://www.world-nuclear-news.org/articles/what-is-the-budget-for-canadas-first-smr-project.

World Nuclear Association (2025b) "United States of America – Reactor Database", [online], Available at: https://world-nuclear.org/nuclear-reactor-database/summary/United%20States%20Of%20America.

Zhong, W. (n.d.). Command injection. [online] OWASP Foundation. Available at: https://owasp.org/www-community/attacks/Command_Injection

Cyber Security in the Subsea Telecommunication Cable Networks

Martti Lehto[1], Petro Julkunen[2] and Hans Hartikainen[1]
[1]Faculty of Information technology, University of Jyväskylä, Finland
[2]Department of Technical Physics, Faculty of Science, Forestry and Technology, University of Eastern Finland, Kuopio, Finland

martti.lehto@jyu.fi
petro.julkunen@uef.fi
hans.k.hartikainen@jyu.fi

Abstract: More than 95% of international internet and telecommunication data is transmitted through subsea fiber-optic cables, valued for their cost-efficiency, low latency, and high capacity. The cables connect the network of data centers around the world to form our internet. This physical infrastructure is critical to the modern world as most modern services depend on it. There are currently over 550 commercial subsea telecommunication cables in-service and many more planned or being constructed. It is crucial to provide resilience of the subsea cables against environmental, human, and other activities that may break any one or multiple cables. Despite their critical importance to global communications, these vast networks remain vulnerable to a range of cyber threats. The remote location of subsea cables does not preclude risks; rather, their endpoints, landing stations, and network management systems are susceptible to malicious actors seeking to intercept, disrupt, or manipulate data flows. As international reliance on these information systems increases, so does the strategic imperative to secure them against espionage, sabotage, and emerging cyber-attacks. Successful cyber-attacks are predicated on exploiting system vulnerabilities. Vulnerability can be defined as exploitable weaknesses or deficiencies in a system, device or its design that allow cyber attackers to execute cyber-attacks. A weakness in system security procedures, software applications, policies and procedures and regulatory compliance may be caused by vulnerabilities. The inherent weakness in the system caused by vulnerability increases the probability of a harmful occurrence or exacerbates its consequences. Vulnerabilities can be divided into those that exist in people's actions, processes in the organizations, and technologies. The paper discusses cyber vulnerabilities, threats and management thereof in the telecommunications network.

Keywords: Subsea cable, Vulnerability, Cyber-attack, Cyber security

1. Introduction

Threat, vulnerability, and risk form an intertwined entirety in the cyber world. Vulnerability can be defined as an exploitable weakness or deficiencies in a system, device or its design that allow threat agents/actors to execute illegal actions in the target system. The paper analyzes vulnerabilities among cable operators and vulnerabilities related to cable technology and IT/OT/SCADA systems (software & hardware).

A cyber threat compiles information related to potential cyber threats targeting various entities such as a system, enterprise, System of Systems (SoS), region, or a critical infrastructure sector. This model serves as a foundational element for diverse tasks across different scopes. Achieving comprehensive cyber security necessitates a thorough analysis of a system of systems (or sub-system) against a spectrum of threat events.

Defining threats within the cyber space realm presents challenges, primarily due to the elusive nature of attack origins, the intricate motives driving them, and the unpredictable unfolding of events. The complexities are further heightened by the blurred boundaries between national, international, public, and private interests, exacerbated by the global nature of cyber threats and the rapid evolution of technology.

Small countries, such as many European nations and island nations that have limited cable access points, are vulnerable to cyber-attacks on the cables, which could have widespread disruptions that go beyond the immediately targeted nation or region as the cables are part of larger networks. Depending on the motivation of the threat actors, the vulnerable points within the cable network may vary. (ODNI, 2017; Gallagher & Carter, 2023)

This paper describes the findings of a research project that examined cyber threats targeting Critical Subsea Infrastructure (CSI). The study focuses on PPT (People-Processes-Technology) vulnerabilities in CSI. The paper outlines the cyber threat environment, identifies vulnerabilities to PPT, and estimates potential impacts. This study consists of conducting a system description, cyber threat analysis, vulnerability analysis, cyber-attack model analysis and impact analysis.

2. System of System Description

A system of systems (SoS) is a collection of systems, each capable of independent operation, that interoperate together to achieve additional desired capabilities. The SoS concept describes the integration of many

independent, self-contained systems to satisfy needs of the organization. The synthesis of these very large systems often results in different problems than those presented by the design of a single, yet complex, system.

2.1 Undersea Cable Network

The architecture of a typical subsea telecommunication cable network is divided into elements of wet plants and dry plants with cable landing stations (CLSs), beach manholes, subsea cables, and repeaters, gain equalizers, amplifiers and branching units. The subsea communications cable connects the landing station which houses the subsea line terminal equipment, where the subsea telecommunication cable is divided into the optical fiber and the power supply line. (Julkunen, 2025)

Amplifiers and repeaters are both components that enhance the signal in order to mitigate the amplitude loss over extended distances. Repeaters are placed "in-line" (placed in intervals along the cable) and amplifiers "off-line" (placed at the end of the cable run). (Chesnoy, 2016)

2.2 Telecommunication Cable

An undersea cable's structure consists of a central bundle of glass optical fibers for data transmission, surrounded by multiple layers of protective materials, including gel, a copper conductor for power, steel strength members for structural integrity, and a robust outer jacket made of polyethylene, nylon, and additional steel wires or aramid yarns. Repeaters are strategically placed along the cable to amplify weak light signals over long distances, and cable landing stations are where the undersea cables connect to terrestrial networks. The constituents of the cable itself, listing from the center outwards, are first, the data-carrying optical fibers and core, strength wires, copper sheath to power the underwater equipment, insulation jacket, followed by varying amounts of protective layers. (Chesnoy, 2016)

2.3 Supervisory Control and Data Acquisition

A key component of critical infrastructure is operational technology (OT), which includes the most important ones: Industrial Control System (ICS), Supervisory Control and Data Acquisition (SCADA), Distributed Control System (DCS) and Programmable Logic Controller (PLC). These systems are key components of the subsea cable network. SCADA consists of physical and software components, which include sensors and actuators, data transmission networks, controllers, communication devices, user interface (HMI) and software. (Lehto, 2022; Hartikainen, 2025)

3. Stakeholders in the Subsea Cable Network

The owners invest in subsea cables/pipelines to increase the capacity for meeting the demand or to expand coverage to serve new regions and customers. Telecommunication cables are often owned and installed by private companies such as telecommunication carriers or subsea cable companies but can also be state-owned. Ownerships can also be distributed instead of a single owner to consortiums or public-private partnerships (Reverdy & Skenderoski, 2015).

Suppliers for subsea telecommunication systems include companies that are fully integrated system suppliers, system integrators, cable suppliers, Subsea Line Terminal Equipment (SLTE) suppliers, and special component suppliers. The suppliers often have access to the cables and pipelines for construction, repairs and maintenance, and possess a deep know-how on the systems as well as proper hardware and the ability to analyze and repair damage.

The stakeholders include governments, organizations and people connected via the subsea cables and pipelines, and potentially those with financial and strategic interests in those connections, regional or global. The operators and stakeholders are intertwined and connected as presented in Figure 1. The connections are generated by the operators' and stakeholders' interdependencies and their resources and assets. (Gallagher & Carter, 2023; Julkunen 2025).

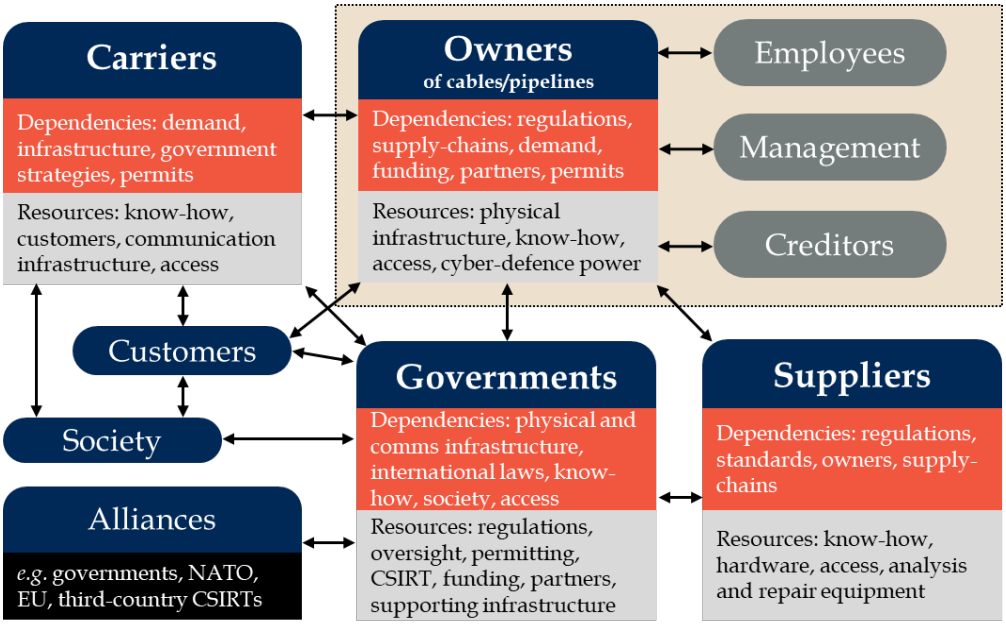


Figure 1: Network of operators and stakeholders

The people of the operator cable/pipeline operator organizations include the employees, management, potential consultants and subcontractors, as well as investors and auditors. These people are considered to have access that exceeds that of a person operating outside of the organization. These people are hence important concerning limiting access to cables and pipelines, and the associated systems including the software and hardware and monitoring systems, like SCADA. Hence, the cable/pipeline organizations and their safety are dependent on the cyber security skills and know-how of the people in those organizations as well as the level of understanding on cyber security. The people in those organizations generate processes initiated by internal or external motivation.

An insider is a cyber-attacker who has legitimate access to the organization and its knowledge and resources. (UK-NCSC, 2022) Also ENISA considers also human errors and carelessness to be included. (ENISA, 2024) While the motivation of insiders can broadly fit in with our wider taxonomy, they are a special category warranting consideration. Empirical results suggest main motivations for insider cyber sabotage to be revenge, personal gain, poor performance and addiction. (Maasberg et al., 2020) Most prevalent insider incidents are fraud (44%), sabotage (24%) and intellectual property theft (16%), with miscellaneous (12%) and multicategory (4%) making up the rest. (Collins, 2013)

4. Threat Analysis

Threats in cyber space are difficult to define due to challenges in defining the source of attacks and the motives that drive them, or even to foresee the course of an attack as it unfolds. The identification of cyber threats is further complicated by the difficulty in defining the boundaries between national, international, public, and private interests. Because threats in cyber space are global and involve rapid technological developments, the struggle to counter them is ever-changing and increasingly complicated. (Lehto, 2022)

For this study, a practical threat taxonomy based on the motivation of the attacker has been developed. The threats included in the suggested threat model are all applicable to the critical infrastructure assets presented in the chapter. The presented threat taxonomy mainly covers cyber security threats; that is, threats applying to the whole OT environment.

A common threat model is a six-fold classification based on motivational factors: cyber vandalism, cybercrime, cyber intelligence, cyber terrorism, cyber sabotage, and cyber warfare. The motivation also affects the attacker's targeting and methods. While a terrorist seeks visibility, a spy wishes to stay unnoticed to gain information. Different attacker archetypes (Table 1) also have different triggers, i.e. events or conditions that cause the attacker to execute the attack.

Table 1: Cyber threat adversary archetypes' attributes

	Vandalism	Crime	Intelligence	Terrorism	Sabotage	Warfare operations
Motivation and goal	Pursuit of changing the environment or cause harm. *Egoism*	Pursuit of financial gain *Financial*	Pursuit of information to support decision-making. *Information*	Pursuit of social instability and influencing political decision-making. *Anarchy*	Pursuit of instability, chaos, political change, infrastructure paralysis. *Paralysis*	Pursuit of paralysis/destruction of nation's digital infrastructure. *Political or military dominance.*
Target	Digital services of the governments, companies, and citizens' information systems.	Digital services of the governments, companies, and citizens' information systems.	Data and information of governments, companies, important individuals	Data and information of governments, companies and critical infrastructure.	Nation's critical infrastructure.	Nation's critical infrastructure (civilian or military).
Attackers' capability	Acquired	Augmented	Advanced	Advanced	Integrated	Integrated
Impacts	Disruptions in digital services	Financial losses	Data and information loss	Disruptions and paralysis in digital services and critical infrastructure	Paralysis and damage in critical infrastructure	Paralysis, damage and destruction in critical infrastructure

5. Vulnerability Analysis

Threat, vulnerability, and risk form an intertwined entirety in the cyber world. Vulnerability can be defined as an exploitable weakness or deficiencies in a system, device or its design that allow cyber-attackers to execute cyber-attacks. Vulnerabilities may be the outcome of a weakness in system security procedures, software applications, policies and procedures and regulatory compliance. Vulnerabilities are inherent weaknesses in the system which increases the probability of an occurrence or exacerbates its consequences. (Lehto, 2022)

Vulnerabilities can be divided into those that exist in people's actions, processes in the organizations, and technologies.

People refer to the human resources available at the firm's disposal. The people are the ones who do the tasks described in the process. Employees are often victims of social engineering tactics and may end up unknowingly providing attackers with login credentials or classified organization data. (Lehto, 2022)

Processes are crucial in defining how the organization's activities, roles and documentation are used to mitigate the risks to the organization's information. Process vulnerabilities among others are lack of written security policy, poor regulating policy, lack of security awareness and training, and poor adherence to security, lack of access control and non-existence of disaster/contingency plan.

Technology solutions protect against cyber risks that may arise from network vulnerabilities, but technology itself contains vulnerabilities (hardware, HW and software, SW). So, technological vulnerabilities are security holes in a system. (Hartikainen, 2025)

Vulnerabilities concerning people and processes in cyber security of critical subsea infrastructure. The operational vulnerabilities, i.e. vulnerabilities associated with human action and processes. The entire vulnerability scene of the cyber space of critical subsea infrastructure has been summarized in Figure 2.

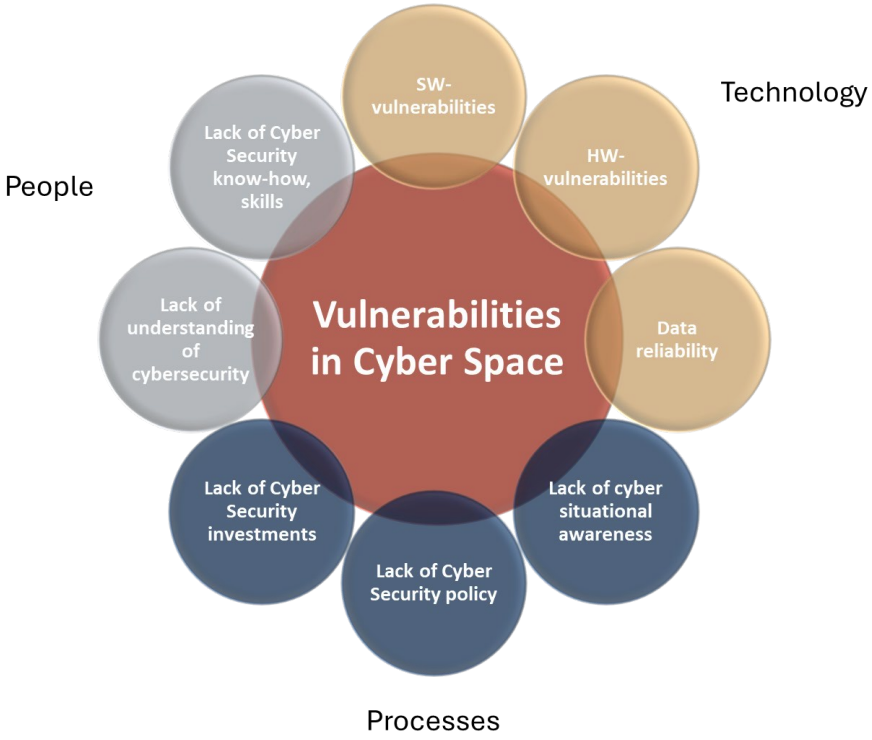


Figure 2: Vulnerabilities associated with cyber space of critical infrastructure

5.1 Vulnerabilities in People

People-related threats are referred to as insiders. These are people who have, or had, authorized access to information systems or information about an organization's information resources. Insider threats can be malicious and unintentional based on their intentions. A malicious insider may exploit their own access to information systems. Motives can include revenge, money, the need for recognition, or other personal reasons. An unintentional insider threat includes cases where an insider causes damage either through their actions or by failing to act. The most significant threat is unwary employees who fall for social engineering tactics like scam emails or phone calls. Social engineering is used by threat actors to take advantage of individuals' vulnerabilities by manipulating people through human interaction to form trust and confidence to compromise sensitive information. It is usually a matter of lack of knowledge and awareness that makes individuals most prone to phishing attacks. One group is made up of employees whose information system has been hacked due to the organization's poor cyber security. In this case, their computer is used to gain access to the organization's information system without their knowledge. (Ihanus, 2023)

Zimmermann and Renaud (2019) reported following vulnerabilities related to people at the individual level: 1) lack of cyber awareness, knowledge and skills, 2) lack of accountability, 3) lack of policies and compliance, 4) not following the best security practice, 5) not sharing responsibility, 6) malicious employees, and 7) cyber criminals (outside the organizations). The lack of cyber security awareness and knowledge may make individuals more prone to social engineering attempts, like phishing, as well as openly sharing information relevant to cyber security. The general issues with noncompliance with cyber security policies are known to be utilized for gaining access (Pham et al., 2017; Victor-Mgbachi, 2024). This connects vulnerabilities in cyber security policies with lack of proper security updates, inappropriate applications installed on critical systems, authentication weaknesses, unauthenticated PLC and remote terminal units (RTU) network connections, remote access supervision and interconnection management (ENISA, 2016).

5.2 Vulnerabilities in Processes

Vulnerabilities in processes include: 1) lack of cyber security know-how and skills, 2) lack of understanding of cyber security, 3) insufficient of cyber security investments, 4) lack of cyber security policy and 5) lack of cyber situational awareness.

The study in the United States identified the following process-related vulnerabilities for subsea telecommunication cables: 1) the private ownership of the cables and cross multiple jurisdictions, 2) the federal

review and permitting process for marine infrastructure may not consider cable protection; 3) lack of a robust protection strategy, 4) lack of a responsible federal agency on cable protection, and 5) gaps in leadership for reporting and investigating unusual activity, and facilitating cable issue resolution (Gallagher & Carter, 2023).

5.3 Vulnerabilities in Technology

Software vulnerability is a bug in program coding, configuration, or management. A program can be an algorithm, application, operating system or browser and control software like communication protocols and drives devices. Hackers exploit vulnerabilities in software attacks to force systems to give them access to unauthorized data, execute malicious code, obtain remote control, or cause the system to spread infections. The CyLab Sustainable Computing Consortium at Carnegie Mellon University estimates that "commercial software has 20-30 code bugs for every 1000 lines of code". Already over two decades ago, Applied Visions, Inc. estimated that 111 billion lines of new software code containing billions of vulnerabilities are coded every year. (Delio, 2004)

Hardware vulnerability is an exploitable weakness in a computer system that enables attack through remote or physical access to system hardware. Hardware vulnerabilities are difficult to identify. In January 2018, the entire computer industry was put on alert by two new processor vulnerabilities dubbed Meltdown and Spectre that defeated the fundamental OS security boundaries separating kernel and user space memory. The flaws stemmed from a performance feature of modern CPUs known as speculative execution. (Constantin, 2021)

IT/OT convergence (sub-aspect of sc. Industry 4.0) is the recent trend of increased connectivity between IT and OT systems, for various economical, strategical and operational reasons. However, this also introduces challenges to cyber security via widening of the attack surface. (ENISA, 2019; NSTAC, 2022)

5.4 SCADA-specific Issues

Nelligere et al. (2023) lists vulnerabilities among different levels, from process control to enterprise. Alanazi, Mahmood & Chowdhury (2023) divides the vulnerabilities into two categories: issues resulting from implementation and issues resulting from configuration. Upadhyay & Sampalli (2020) have these two analogous categories along with network security and protocol vulnerabilities.

Implementation

Improper validation of received input can lead to unintended privileges or otherwise unintended actions. Techniques might involve e.g. command or SQL injection, path traversal, or buffer overflow. Other failures to manage memory buffers can also result in further access or destruction of data. Control flow issues such as race conditions can also cause similar effects. (Upadhyay & Sampalli, 2020, Alanazi, Mahmood & Chowdhury, 2023)

Many of these issues are exacerbated by reliance on legacy protocols and devices, remote installations (this particularly applies subsea) and lack of knowledgeable people about particular of said systems. (Nazir, Patel & Patel, 2017, Nelligere et al., 2023)

Network and protocols

General principles of cyber security to great extent apply. Lack or improper configuration of firewalls/IDS/SIEM, network segmentation, access control lists and encryption are common issues. (Nelligere et al., 2023)

Proper encryption is not always feasible due to proprietary or legacy protocols involved, for e.g. MODBUS. In general, traffic sniffing, IP spoofing and Man-in-the-Middle attacks are among relevant vulnerabilities. (Upadhyay & Sampalli, 2020, Alanazi, Mahmood & Chowdhury, 2023)

Configuration

Weak password requirements, hard-coded credentials or other default passwords can enable the attacker to initially access or move around within the system. Also lack of encryption as mentioned previously can lead to exposed credentials. (Upadhyay & Sampalli 2020, Nelligere et al. 2023)

Deficiencies and faults in access control and authentication can also lead to undesired access to target network. Examples include (but are not limited to) bypassing authentication altogether, replay attacks and brute force attacks. (Alanazi, Mahmood & Chowdhury, 2023)

6. Impact Analysis

Impacts are the end point impacts that manifest after successful attack paths are completed, and they are agnostic to the techniques and technologies used. The effects of a cyber-attack can be diverse, such as: Account Access Removal, Damage to Property, Data Destruction, Data Encrypted, Data Manipulation, Web Page Defacement, Denial of System Control, Denial of Service, Disk Wipe, Firmware Corruption, Inhibit System Recovery, Loss of Availability, Loss of System Control, Loss of Productivity and Revenue, Manipulation of Control and View, Network Denial of Service, Resource Hijacking, Service Stop, System Shutdown/Reboot, Theft of Data and Information. (MITRE, 2025)

Impact propagation from affected systems to others is a risk in connected environment. Considering all the possibilities for attacks, it is resource intensive. Behind an attack there are varying triggers depending on the attacker archetype. One way to do an impact analysis is to create scenarios appropriate to the situation. Plausible scenarios can tell a story of what might happen. They give material for creating test cases and tabletop exercises, but do not comprehensively present everything that can happen. The more complicated scenarios aid in revealing dependencies between systems and events.

Cyber-attack models outline the structure of attack-vectors. Threat modelling abstracts a system, profiles potential attackers along with their goals and methods, and enumerates possible threats. By mapping potential attack paths and analyzing vulnerabilities through frameworks like MITRE ATT&CK, security analysts can understand attacker behavior, prioritize defenses, and reduce risk. (MITRE, 2025)

The study developed 14 scenarios for various cyber-attacks summarized in Table 2. The factors were initial access, vulnerabilities exploited, execution, attacker motivation, impact tactics, likelihood of attack and impact. The scenarios were ranked based on their individual estimated risks (impact x likelihood). (NIST, 2012) The overall risks of the scenarios presented are considered low or high. This overall risk represents a view that considers the target of the attack as well as the primary stakeholders. In this view, the highest risks were observed with scenario related to ransomware attack, and scenario related to exploitation of vendor-caused vulnerabilities. Considering the scenarios, we considered the top six most relevant scenarios, i.e. the ones that rank HIGH in risk. Out of these, we discerned the most utilized vulnerabilities. Table 3 illustrates most often occurring vulnerability categories and super categories in top 6 most relevant scenarios.

Table 2: Summary of developed scenarios

Scenario	Impact	Likelihood	Risk
Exploiting vendor vulnerabilities	HIGH	HIGH	HIGH
Malicious insider activities	MODERATE	HIGH	HIGH
Espionage on cable operators	HIGH	MODERATE	HIGH
Injecting false sensor information	MODERATE	HIGH	HIGH
Ransomware attack	HIGH	MODERATE	HIGH
Distributed denial-of-service attack	HIGH	MODERATE	HIGH
Phishing attack	MODERATE	MODERATE	MODERATE
Supply-chain attack	MODERATE	MODERATE	MODERATE
Hybrid attack (to energy network)	MODERATE	MODERATE	MODERATE
Private cable operator conducting espionage	HIGH	LOW	MODERATE
Misconfiguration of firewall	MODERATE	MODERATE	MODERATE
Credential exposure via misconfigured cloud	LOW	MODERATE	MODERATE
GPS spoofing	MODERATE	MODERATE	MODERATE
Fiber tapping telecommunication cable	LOW	LOW	LOW

Table 3: Most often occurring vulnerabilities

Vulnerability	Occurrences in scenarios
Lack of cyber security policy (super category)	6
Lack of understanding of cyber security (super category)	4

Vulnerability	Occurrences in scenarios
Poor cyber hygiene	4
Lack of security know-how and skills (super category)	3
General compliance issues	3
Lack of a robust protection strategy	3
Cyber security awareness is not a priority	2
Lack of preparedness for blackouts and internet outage	2
Dependencies on suppliers and service providers	2
Situational awareness not a priority	2
Insufficient cyber security investments	2
Low number of cable / pipeline access points / country	2
Insider threat	2
Operations divided and access granted to multiple operators	2
Lack of good practices in supply-chains	1
Lack of trained personnel	1
Private ownership of the cables	1

On a super category level, lack of cyber security policy was discovered to be a common factor in all the top 6 scenarios, with a closely related lack of understanding of Cyber security trailing closely behind with 4 scenarios, while rest of the super categories only concern 2 or 3 of the attacks. In addition, many of the most often occurring sub issues are contained within this, and thus formulation and upkeep of a quality cyber security policy is a natural recommendation for prioritization, if not in order already.

For example, one of the most significant initial attack vectors is phishing, thus having the personnel trained on the fundamental basics and up-to-date on the newest types of phishing attempts is crucial; this was singled out as one of the simple key issues in private discussions with a Digital Forensics and Incident Response (DFIR) operator. Training of personnel is key part of a cyber security policy.

Vendor-caused vulnerabilities scenario ranked HIGH both on likelihood and impact. Having a strategy in place for identifying and handling dependencies on suppliers and service providers is key in mitigating vendor-related vulnerabilities, both with respect to software and hardware, i.e. adopting a C-SCRM program (Cyber security Supply Chain Risk Management). This applies for both the external vendor-provided devices and software. According NIST Cyber security Framework 2.0: Quick-Start Guide for Cyber security Supply Chain Risk Management (C-SCRM) (NIST 2024; NIST, 2022):

- Having a Supply Chain Risk Management strategy (along with objectives, policies and processes) in place.
- List external technology suppliers and determine the criticality of each.
- Formulate relevant roles and requirements and communicate them to partners and stakeholders.

Where legacy systems cannot be updated or replaced, and as good practice in general, isolating the potentially vulnerable OT systems by network segmentation is crucial (as in the typical Purdue models). (Williams, 1992)

7. Discussion and Conclusion

Telecommunications and electricity sectors are considered critical and dependent on each other, hence possessing risks that an incident in one sector would affect the other (NIS Cooperation Group, 2024). All other sectors considered critical by European Union (2022a) are dependent on them. After identification of such vulnerabilities based on previous governmental, organizational and scientific reports, scenario-based modelling was conducted to consider various possible incidents that originate from cyber threats related to CSI. Based on the scenarios a rough risk assessment was made following common practices (NIST, 2012; ENISA, 2022). The risk assessment indicated that the overall risks caused by potential cyber threats are moderate to high with the greatest risks focused on the target organization. Cyber threats combined with vulnerabilities comprise cyber risks. In general, it is considered that cyber threats exist, and we cannot influence them, while we can reduce vulnerabilities to a point hence enabling us to manage the risks potentially to an acceptable level.

The study found that critical submarine infrastructure may be particularly vulnerable to cyberattacks due to organizational cybersecurity deficiencies. Potential key vulnerabilities include limited cybersecurity skills, lack of awareness, insufficient investment, fragmented policies, inadequate monitoring and shortcomings in technology. Additional concerns arise from private ownership and multinational supply chains. These conditions may complicate coordinated protection and risk management efforts across jurisdictions.

The research emphasizes the requirement for improved cyber security awareness, harmonized policy implementation, increased investment in training, and enhanced protection strategies, some of which have been recently implemented at the regulatory level in the EU but need to be brought to the organizational level.

It is expected that the resilience to cyber threats is increasing through investments and awareness training as well as the protective methods described above. This will help control the impact of the threat, as the likelihood of threats is on the rise. According to International Institute for Strategic Studies (Bentham, 2025), the vulnerabilities in telecommunication cables are increasing, as supposedly the attacks have become more feasible for smaller nations and others than state-sponsored actors. Hence, the mitigated risks are likely to remain balanced by the increase in likelihood and decrease in impact. In other words, the threat keeps rising, but we can perhaps reduce the vulnerabilities exploited.

Ensuring the integrity and resilience of subsea cable networks demands a blend of robust encryption, vigilant monitoring, international cooperation, cyber security policies and technical safeguards - measures essential to defending the digital world together.

Ethics declaration: An ethics declaration was not required for the research.
AI declaration: AI tools were not used in the creation of this paper.

References

Alanazi M., Mahmood A. and Chowdhury M. 2023. SCADA vulnerabilities and attacks: A review of the state-of-the-art and open issues, Computers & security. https://doi.org/10.1016/j.cose.2022.103028

Bentham J. 2025. Subsea advances and challenges for the Asia-Pacific, retrieved 18.5.2025, https://www.iiss.org/online-analysis/online-analysis/2025/05/subsea-advances-and-challenges-for-the-asia-pacific/

Chesnoy J. 2016. Undersea Fiber Communication Systems. Academic Press.

Collins M. 2013. Analyzing Insider Threat Data in the MERIT Database, CERT Insider Threat Center, retrieved 15.1.2025. https://insights.sei.cmu.edu/insider-threat/2013/10/-analyzing-insider-threat-data-in-the-merit-database.html

Constantin L. 2021. 33 hardware and firmware vulnerabilities: A guide to the threats. retrieved 10.4.2025 https://www.csoonline.com/article/3410046/hardware-and-firmware-vulnerabilities-a-guide-to-the-threats.html

Delio M. 2004. Linux: Fewer Bugs Than Rivals, Wired blog, Dec 14, 2004, retrieved 25.9.2025. https://www.wired.com/2004/12/linux-fewer-bugs-than-rivals/

ENISA 2016. Communication network dependencies for ICS/SCADA Systems.

ENISA 2019. Industry 4.0 - Cybersecurity Challenges and Recommendations.

ENISA 2022. Interoperable EU Risk Management Framework.

ENISA 2024. ENISA Threat Landscape 2024.

Gallagher J. C. & Carter N. T. 2023. Protection of Undersea Telecommunication Cables: Issues for Congress. Congressional Research Service. R47648.

Hartikainen H. 2025. Technical cyber threats in undersea infrastructure, University of Jyväskylä, master's thesis.

Ihanus J. 2023. Sisäpiiriuhkan hyökkäysketju – sisäpiiriuhkan hyökkäysvaiheiden mallintaminen, University of Jyväskylä, Master Thesis.

Julkunen P. 2025. Cyber threat analysis of people and organizational processes in critical submarine infrastructure, University of Jyväskylä, master's thesis.

Lehto M. 2022. Cyber-attacks Against Critical Infrastructure, in Lehto M. and Neittaanmäki P. (Eds.) Cyber Security: Critical Infrastructure Protection, in series Computation Methods in Applied Sciences, Springer 2022, pages 3-42. ISBN: 978-3-030-91293-2.

Maasberg M., Zhang X., Ko M., Miller S. R., & Beebe N. L. 2020. An Analysis of Motive and Observable Behavioral Indicators Associated With Insider Cyber-Sabotage and Other Attacks. IEEE engineering management review, 48(2), 151-165. https://doi.org/10.1109/EMR.2020.2989108

MITRE 2025. MITRE ATT&CK ICS dataset. Accessed 5 May 2025, https://attack.mitre.org/resources/attack-data-and-tools/

Nazir S., Patel S., & Patel D. 2017. Assessing and augmenting SCADA cyber security: A survey of techniques. Computers & security, 70, 436-454.

Nelligere S. L., Swamy G. T., Akheel M. A., & Chandramma R. 2023. Vulnerability Assessment and Analysis of SCADA and Foundation Fieldbus on Industrial Control System (ICS) Networks: A Literature Review. ICFAI journal of computer sciences, 17(2), 34-65.

NIS Cooperation Group 2024. EU cybersecurity risk evaluation and scenarios for the telecommunications and electricity sectors. European Commission.

NIST 2012. Special Publication 800-30 - Guide for Conducting Risk Assessments. U.S. Department of Commerce.

NIST 2022. Cybersecurity supply chain risk management practices for systems and organizations. U.S. Department of Commerce. NIST Special Publication 800-161r1

NIST 2024. NIST Cybersecurity Framework 2.0: Quick-Start Guide for Cybersecurity Supply Chain Risk Management (C-SCRM). NIST Special Publication No. 1305

NSTAC 2022. Information Technology and Operational Technology Convergence, President's National Security Telecommunications Advisory Committee.

ODNI 2017. Threats to Undersea Cable Communications. Office of the Director of National Intelligence.

Pham H.C., Brennan L. and Richardson J. 2017. Review of behavioural theories in security compliance and research challenges. Informing Science + IT Education Conferences, Ho Chi Minh City, Vietnam.

Reverdy D. & Skenderoski, I. 2015. Submarine Cables: Structuring and Financing Options.

UK-NCSC 2022. Reducing data exfiltration by malicious insiders, United Kingdom National Cyber Security Centre, https://www.ncsc.gov.uk/guidance/reducing-data-exfiltration-by-malicious-insiders

Upadhyay D., & Sampalli S. 2020. SCADA (Supervisory Control and Data Acquisition) systems: Vulnerability assessment and security recommendations. Computers & security, 89, 101666.

Victor-Mgbachi T. 2024. Navigating Cybersecurity Beyond Compliance: Understanding Your Threat Landscape and Vulnerabilities. IRE Journals 7(7): 70-81.

Williams T. J. 1992. The Purdue enterprise reference architecture: a technical guide for CIM planning and implementation. Research Triangle Park, NC: Instrument Society of America.

Zimmermann V. & Renaud K. 2019. Moving from a 'human-as-problem" to a 'human-as-solution" cybersecurity mindset. International Journal of Human-Computer Studies 131: 169-187.

On the Establishment of Trust: Challenges, Opportunities and Socio-Cultural Factors

Christoph Lipps[1], Denise Scharwatz[2] and Henry Collier[3]
[1]Intelligent Networks Research Group, German Research Center for Artificial Intelligence, Kaiserslautern, Germany
[2] RPTU University Kaiserslautern-Landau, Kaiserslautern, Germany
[3] Thomas Edison State University, Trenton, USA

Christoph.Lipps@dfki.de
D.Scharwatz@rptu.de
hcollier@tesu.edu

Abstract: Trust is one of the fundamental necessities of human beings and, according to *Stephen R. Covey*, not simply the "glue of life", but also the "most essential ingredient in effective communication". Even though the principles and importance of trust are as old as humanity itself -in ancient times, trusting strangers could mean the difference between life and death, and thus pose an immediate threat to one's social tribe- trust is gaining increasing attention, particularly in light of tomorrow's all-electric society and the decentralization and globalization associated with it. Contrary to previous decades, physical proximity is no longer necessary to access a system; access is possible (almost) anytime, (almost) anywhere. However, trust is a multidimensional concept depending on a multitude of aspects, such as the specific application, the value of the resource, and the available technology, but also -for instance- on the people who are managing access to systems and applications. As societies become increasingly aware of data breaches, algorithmic surveillance, and commercialization of personal data, cultural values have shifted toward individual agency and distrust of centralized institutions. As trust could never be taken for granted it rather must be earned, orchestrated and continuously verified. Technologies like multi-factor authentication (MFA) reflect this, offering multi-layered mechanisms for identity verification and access protection. These developments illustrate a broader societal reshaping of trust -from implicit trust toward conditional, data-driven security- and show how technological design is shaped by evolving social anxieties and expectations. Against this background, this work focuses on socio-cultural influences on the definition of trust and trustworthiness, such as origin (geography, culture, political system) or educational background and training. In particular, it follows research questions including: i) How do socio-cultural perceptions of trust and identity shape the development and adoption of digital security technologies such as MFA; ii) in what/which ways does the rise of digital surveillance and data breaches influence societal expectations of privacy and trust in technological systems?; and iii) how is the concept of trust redefined in the digital age, and what role do authentication technologies play in mediating this transformation?

Keywords: Trust**,** Trustworthiness, Critical infrastructure security, Cyber security and resilience

1. Why Trust Matters: Geopolitical Drivers for Trust and Trustworthiness in Cybersecurity

The conceptual foundations of trust in digital systems are undergoing profound transformation. While *trust* traditionally refers to a subjective attitude toward people, systems, or institutions, *trustworthiness* denotes the objective characteristics and behaviors that render an entity reliable (Fettweis, et al., 2025). This distinction has become vital in a digital society where interactions are increasingly mediated by complex technical infrastructures rather than direct human relationships. Trustworthiness must therefore be established, demonstrated, and continuously verified. Security mechanisms such as multi-factor authentication (MFA) illustrate this evolution: they do not create trust themselves, but provide measurable, verifiable properties -such as strong identity assurance- strengthening the perception of trustworthiness. As digital infrastructures become more interconnected and exposed to sophisticated threats, the capability to design and evaluate trustworthiness has emerged as a central prerequisite for secure communication and resilient critical infrastructures (Sithungu & Lipps, 2025).

Building on this distinction, trust can be understood -following (Luhmann, 2017)- as a subjective expectation that enables individuals to navigate complexity by relying on assumptions about the behavior of others or of systems. It is shaped by prior experience, perceived risk, and contextual interpretation. Trustworthiness, by contrast, refers to the objectively assessable qualities of a system or entity, expressed through measurable attributes such as reliability, security assurances, or compliance with standards like those defined by National Institute of Standards and Technology (NIST) -e.g. SP 800-160, SP 800-53, SP 800-12-, and International Organization for Standardization (ISO), -e.g. ISO/IEC TS 5723:2022. Although these two concepts are often implicitly treated as mutually reinforcing, their relationship is neither linear nor guaranteed. A system that is demonstrably more trustworthy in a technical sense does not necessarily evoke higher levels of human trust, particularly when cognitive biases, sociocultural factors, or limited transparency hinder users' ability to interpret

or appreciate such assurances. This conceptual tension is increasingly relevant in digital environments where subjective trust judgments depend heavily on abstract indicators of technical trustworthiness rather than direct interpersonal experience.

Against this backdrop, trust and trustworthiness represent two of the most strategically important -yet increasingly fragile- pillars of modern cybersecurity. Rising geopolitical tensions and accelerating digital interdependence challenge the ability of states, organizations, and operators of critical infrastructure to build and maintain trustworthy technological ecosystems. Recent incidents illustrate the erosion of this trustworthiness: the compromise of a Norwegian dam's control system in April 2025 through weak access credentials (Pritchard, 2026); the "Salt Typhoon" attacks on U.S. telecommunications providers, which targeted identity verification and secure communication systems (Boston Institute of Analytics, 2025); and the ransomware attack on Colonial Pipeline in 2021, one of the most impactful cyberattacks on U.S. energy infrastructure (Lipps et al., 2022). These cases demonstrate how even seemingly isolated intrusions can reveal systemic weaknesses, marking a broader trend in which cyber operations increasingly target the integrity, availability, and reliability of essential services rather than merely extracting data.

At the same time, the global threat landscape is shifting rapidly. The rise of generative AI (GenAI) has changed the nature of cyberattacks: attackers now automate intrusion attempts, craft convincing social engineering campaigns, and conduct long-term operations designed to undermine institutional trust rather than simply compromise systems. Identity-based intrusions exploiting valid credentials, trusted AI agents, or federated identity platforms are proliferating, eroding traditional assumptions about trust boundaries and exposing the limitations of perimeter-based security models.

Critical infrastructures -energy grids, water treatment facilities, transportation networks, and healthcare systems- are particularly vulnerable. Many rely on outdated operational technologies (OT) never designed for secure connectivity. Regulatory responses such as the NIS2 Directive (NIS 2 Directive, 2022) and the Cyber Resilience Act (Cyber Resilience Act, 2024) reflect the urgency of strengthening systemic trustworthiness, but compliance alone cannot guarantee resilience. Instead, deeper insight is required into how trust and trustworthiness can be established and maintained across distributed, heterogeneous, and geopolitically sensitive environments.

Simultaneously, emerging technologies create new trust dependencies. AI-enabled security systems offer improved detection capabilities but introduce opaque decision-making processes and new attack surfaces (Chokkanathan et al., 2024). Research on AI-driven Zero Trust architectures highlights the necessity of continuous verification, identity assurance, and adaptive trust assessment—principles especially relevant as adversaries become more dynamic. Likewise, the global transition toward postquantum cryptography reflects concerns that quantum capable adversaries could compromise today's cryptographic foundations.

The interplay of these developments -geopolitical fragmentation, AI-assisted cyber operations, vulnerable infrastructure, and the shift toward quantum-resistant security-, creates an urgent need for updated frameworks to strengthen trustworthiness. Such frameworks must be resilient to deception, interoperable across jurisdictions, and capable of functioning even when traditional trust anchors, such as state cooperation or vendor reliability, are destabilized. Building trustworthiness is therefore not merely a technical task, but a strategic requirement for societal stability, economic continuity, and national security.

Accordingly, this work examines how trust and trustworthiness can be systematically conceptualized, built, and maintained in modern communications infrastructures and critical systems. Integrating insights from cybersecurity engineering, policy studies, and emerging-technology research, it aims to provide a comprehensive foundation for designing trust mechanisms capable of withstanding the pressures of contemporary geopolitical and technological realities.

2. Definition of Trust and Trustworthiness

Trust and trustworthiness are often conflated, yet they refer to fundamentally different concepts. Thus, there are very diverse perspectives on the concept of trust, as it

- is a belief, or more importantly a psychological state where someone believes in the integrity, reliability or ability of someone else (TrustTalk, 2026),
- results in one person's willingness to rely upon the other person,
- is a foundational component of personal, business, and societal relationships, to include the relationships within oneself, and

- is a crucial component of life, especially one's social life (Yamamoto, 1990).

So, trust has several components that are interconnected. These components include reliability, honesty, competence, benevolence and integrity. As humans, we are taught to trust other people, to trust situations and to trust that good will prevail over evil. However, from a cybersecurity perspective, trust is most elusive, and misguided trust leads to security incidents.

Trust can be further categorized into three categories: i) logical trust, ii) emotional trust and iii) relationship trust. Logical trust and emotional trust are considered individual trust attributes, while relationship trust is its own category called relational trust attribute (Cho, et al., 2015). Whereas trustworthiness is a quality or characteristic that makes someone worthy of trust. Trust is something that is built based on information from different sources to include a person's reputation, past behaviour, identity, communication and social norms. For example, a religious leader is typically considered to be someone who can be trusted, this is an example of past behaviour, identity and social norms. As a result, a religious leader is considered trustworthy. Trustworthiness is a primary factor in developing and maintaining trust and trusting. Trust and trustworthiness are influential factors impacting effective communication, cooperation, and decision-making (TrustTalk, 2026).

Who we trust and why we trust depends on what kind of person we are as an individual. There is no cookie cutter concept that applies to everyone. At its core, trust is unequivocably influenced by culture. Culture is one of the strongest influences on decision-making. Culture comes in many forms, two of the primary forms are individualism and collectivism. In both forms of culture, the decision-making process is influenced differently. In individualism, the cultural influences tend to lead people to make decisions that result in personal benefit while in the collectivism form, the result is decisions that group benefit. Within each of these forms, there are many sub forms of culture as well, resulting in varied outcomes from within one primary form of culture (Glazer & Karpati, 2014) (Yates & De Oliveira, 2016). A clear example of how culture influences trust and trustworthiness can be found in how people use cultural experiences to frame an individual's understanding of people and situations. Since trust is a psychological state of mind and trustworthiness is quality that makes someone worthy of trust, both are based on a person's experience with someone or a situation (Glazer & Karpati, 2014).

In the world of information technology and cybersecurity, trust and trustworthiness are constantly being evaluated and critiqued. In modern cybersecurity, the community increasingly adopts what is known as a *zero-trust* perspective. While this approach is often described -somewhat misleadingly- as trusting "only what can be verified," it is important to clarify that absolute verification is neither possible nor part of the zero-trust paradigm. The term "zero trust" is a marketing shorthand rather than a literal description: the underlying mechanisms do not eliminate trust but rather aim to minimize the trust base and make trust assumptions explicit, for example by shifting reliance from hardware trust anchors to mathematically verifiable properties. Nonetheless, a residual set of unverified assumptions remains unavoidable, and their extent is not always fully knowable.

This becomes clear when considering practical scenarios. For instance, a recruiter at a leading company must process applications from individuals they have never met. Their role requires opening résumés or portfolios in formats such as PDF or Word -files that inherently carry some risk because they may contain malicious content (Naprys, 2025). Zero-trust concepts can help reduce exposure, but they cannot eliminate underlying trust dependencies entirely. The aim of this work is not to propose a solution to this scenario but to highlight how such challenges reveal ongoing tensions in cybersecurity, especially concerning how the community conceptualizes trust and trustworthiness.

Even in a zero-trust environment, breaches remain possible -not only through human actions, but more generally through the breakdown or violation of underlying trust assumptions, many of which may be implicit or unknown. Social engineering illustrates this challenge: it manipulates individuals into taking actions they otherwise would not, thereby exploiting these residual trust dependencies. Social engineers use all the tools in the book, to include manipulation and deception resulting in attacks that use a person's behaviours and emotions against them to succeed. Social engineers certainly use trust as a weapon, and they are very successful at it. Trust, curiosity, fear and urgency are key components in social engineering techniques that are successful (Pujari & Hussain, 2024). How do social engineers develop trust with a victim, resulting in them being trustworthy in the eyes of the victim. First and foremost, they learn what motivates the person. They then use this information to develop a convincing scam that pushes the person's buttons, using fear and urgency as primary factors.

People become victims of social engineering because their decision-making process has been manipulated. As already mentioned, trust is a key component of the decision-making process. If you trust someone, or something, you are more likely to click on a link or provide your user credentials. This is why many social engineering attacks appear to be people of authority like the Chief Executive Officer (CEO) or the IT department. We are taught to inherently trust these individuals because their position makes them trustworthy and humans are often lazy and do not verify that the email is actually from the person it says it is from.

Trust and trustworthiness are complicated and have been studied many times in various fields (Cho, et al., 2015). All three of the previously noted categories of trust (logical, emotional and relationship) are used by social engineers to conduct their nefarious activities and social engineers implement this attack on trust without care, concern or compassion. Their only goal is to successfully obtain what they want, usually money.

People being vulnerable to social engineer attacks, like phishing, is the direct result of cognitive biases in the decision-making process, for example, inherently trusting that people are good (Pujari & Hussain, 2024). These biases are the result of many social and cultural influences, including religion and social engagement that directly impact trust and trustworthiness.

These dynamics illustrate that trust and trustworthiness operate simultaneously at the human, cultural, and organizational levels. To understand how these concepts manifest in modern digital systems, it is necessary to examine their technological foundations as well as the broader socio-economic forces that shape how trust is created, maintained, and challenged in contemporary society.

3. Technological Foundations of Trust and Trustworthiness

But, things are somewhat different in digital and technical ecosystems, where trustworthiness is not an abstract concept, but a feature established and determined by verifiable and quantifiable mechanisms. Whereas, as mentioned above, human trust is primarily based on perception and experience, technological trustworthiness depends on objective characteristics that can be measured, validated, continuously revalidated, and monitored.

These aspects can be divided into four major components, which are discussed in more detail below and which operationalize different facets of trust and trustworthiness in modern infrastructures: i) the conceptual foundations of technological trust, ii) identity and access management, iii) cryptographic security, and iv) trusted execution environments (TEEs). Whereas it should be emphasized that the term “trusted” in TEE does not imply inherent trustworthiness. In the terminology of trusted computing, TEEs belong to the Trusted Computing Base (TCB), components that must be trusted because they are capable of violating the security policy, not because their behavior can be fully verified. TEEs thus exemplify the unavoidable trust assumptions that persist even in systems designed to maximize verifiable trustworthiness (Lipps, et al. 2018).

At its core, technological trust relies on a set of security and protection goals that ensure systems behave reliably and predictably. While confidentiality, integrity, and availability (CIA) (Schneier, 1995) form a foundational subset -ensuring that data remains private, unaltered, and accessible when needed- these represent only part of the broader landscape. Depending on the specific use case, additional properties such as accountability, authenticity, anonymity, unlinkability, and non-repudiation may be equally essential, as trust cannot be established without the relevant combination of these conditions. However, trustworthiness extends beyond these fundamentals. It encompasses transparency, predictability, and resilience under stress (Van der Ham, 2021). Systems have to prove that they actually deliver defined and warranted performance, are resistant to manipulation (and can detect attempts at manipulation), and can be recovered after failures. This shift from implicit to explicit, evidence-based trust is crucial in an age where interactions take place via complex, distributed infrastructures rather than direct human relationships. Trustworthiness thus becomes a function of verifiable characteristics: secure configurations, verifiable processes, and continuous compliance with defined security principles.

Another essential element is *identity assurance* -that is, the reliable binding of actions to an authenticated principal when accountability is required- because, without it, many security controls become fragile; however, identity in this sense does not preclude anonymity or pseudonymity, which can provide strong security guarantees in use cases (e.g., voting) where linkability to a real-world person is neither necessary nor desirable (Finney & Kindervag, 2023).

Modern identity and access management (IAM) systems implement MFA schemes to reduce reliance on single sign-on credentials by combining knowledge factors (passwords), possession factors (tokens), and inherence factors (biometrics) (Lipps, Herbst, and Schotten, 2021). Beyond MFA, adaptive authentication evaluates

contextual signals -device status, geolocation, behavioral patterns- to dynamically adjust the level of trust. Federated identity systems (e.g., Security Assertion Markup Language (SAML), OAuth 2.0, OpenID Connect) enable cross-domain interoperability but introduce new trust dependencies that require robust governance and cryptographic safeguards for token issuance and revocation. Besides, privileged access management (PAM) adds another layer by controlling high-risk accounts through vaulting, session recording, and just-in-time elevation. Furthermore, Zero Trust enforces continuous authentication and continuous authorization, which are related but not identical. A request is re-authenticated using up-to-date identity and posture signals, and separately re-authorized against policy and risk -even in designs that permit authorization without revealing a civil identity (e.g., pseudonymous credentials). Identity signals become dynamic inputs to authorization, so trust is continually re-evaluated rather than granted by static entitlements (Gambo & Almulhem, 2025).

The technological backbone of this trust lies in the use of cryptographic methods, which provide *conditional* mathematical assurances based on widely accepted computational hardness assumptions. While only very few cryptographic constructs offer formally proven guarantees, modern cryptography enables strong practical protections: encryption schemes preserve confidentiality under assumed adversary limits, and digital signatures or message authentication codes (MACs) provide assurance of integrity and origin. In this sense, cryptography does not eliminate uncertainty, but it offers rigorously analyzed mechanisms that reduce reliance on blind trust by grounding security properties in well-understood mathematical assumptions. Public key infrastructure (PKI) extends these safeguards to distributed systems, enabling secure key exchange and certificate-based chains of trust. However, cryptographic trust is only as strong as its key management practices: secure generation, storage, rotation, and revocation are essential to prevent silent compromises. New threats from quantum computers challenge traditional algorithms such as RSA and ECC, requiring the development of post-quantum cryptography (PQC). Hybrid cryptographic methods that pair classical and post-quantum algorithms are increasingly recommended -not because they ensure forward security, but because they reduce the risk of regression during the transition. Such hybrid constructions require an adversary to compromise both the classical and PQC algorithms, thereby avoiding a scenario where reliance on a single primitive (classical or PQC) could be invalidated by unforeseen cryptanalytic advances. These measures demonstrate that cryptographic trust is not static but must evolve with the capabilities of attackers and be continuously strengthened (Lipps et al. 2020).

While cryptography safeguards data during transmission and in storage, Trusted Execution Environments protect data during use (Shepherd, C., and Markantonakis, 2024). TEEs create isolated enclaves within processors where sensitive computations take place, shielded from the operating system and other applications. This isolation prevents attackers -even those with root access- from manipulating or observing critical processes. TEEs also enable remote verification, allowing systems to prove to external auditors that they are executing approved code in a secure state. This capability is critical for distributed infrastructures where trust must exist across organizational and geographic boundaries. In cloud and edge computing, TEEs support confidential computing and ensure that workloads remain trustworthy even in multi-tenant environments. By anchoring trust in the hardware, TEEs provide a foundation for higher levels of security and complement cryptographic and identity-based controls.

4. The Socio-economic Aspects of Trust and Trustworthiness

Trust and trustworthiness in digital systems emerge not only from technical assurances but also from socio-economic contexts that shape how users perceive, interpret, and act upon those assurances. While engineers ground trustworthiness in measurable properties -as discussed above- users rely on heuristics influenced by cultural norms, education, institutional credibility, and experiences with misinformation. This creates a gap between *engineered trustworthiness* and *perceived trust*, making it essential to understand the socio-economic factors that mediate both.

Empirical evidence shows how strongly social interpretation influences trust. A mixed-methods study conducted in the United Kingdom with 451 respondents demonstrates that many individuals attribute higher trust to AI systems than to humans, largely because AI is perceived as less emotional and less susceptible to manipulation (Gerlich, 2024). This illustrates how users may interpret algorithmic systems through socio-emotional expectations rather than technical guarantees. Conversely, negative experiences can rapidly erode trust. Furthermore, a major U.S. study of 1,045 teens and parents highlights that exposure to deepfakes and fabricated content reduces confidence in information quality, often prompting broader distrust in digital platforms and technologies because users begin doubting the authenticity and provenance of online material (Baron et al., 2025). In both cases, perceptions -not technical properties- drive trust outcomes.

Besides, education and digital literacy shape how people assess trust in uncertain situations. Users with higher media literacy are more capable of distinguishing authentic content from manipulated content, reducing both naïve over trust and unwarranted skepticism. However, most adolescents struggle to determine when content is AI generated, as shown in the same U.S. study material (Baron et al., 2025). This suggests that without adequate training, users face an increasing verification burden they cannot manage alone, especially as generative AI accelerates content production.

Cultural and political contexts further moderate trust: Social norms may emphasize personal autonomy, communal values, or institutional authority, all of which influence how users interpret technological trustworthiness. For example, individuals embedded in environments that value impartiality may more readily accept AI systems as "neutral," whereas individuals in contexts characterized by skepticism toward centralized institutions may distrust AI due to broader political narratives. These orientations help explain why trust levels differ sharply across countries and demographic groups.

Direct interaction with AI systems is another affecting factor of trust, as research shows that users do not necessary require sth. like chatbots to behave human-like and thereby establish trust, but instead, creating interfaces that convey *social presence* -that is, a sense of being emotionally responsive and socially attuned- is more effective than anthropomorphic mimicry (Huynh & Aichner, 2025). This nuance is vital: trust is shaped by *interactional cues*, not by surface-level human imitation. Social presence reduces uncertainty by making AI feel embedded in a familiar relational context, thus reinforcing perceived trustworthiness.

Sector-specific dynamics highlight further socio-economic factors. In healthcare, scenario-based experiments with medical students and physicians demonstrate that explainability and deep integration of AI into clinical workflows significantly increase trust in AI-based clinical decision support systems (AI-CDSS). However, these same features may also heighten professional identity threats, as clinicians worry about losing autonomy or expertise (Ackerhans et al., 2025). Interestingly, the study shows that trust in the system can partially mitigate identity threat, yet accountability mechanisms -such as requiring physician signatures on AI-generated recommendations- can intensify it (Ackerhans et al., 2025). This reveals a complex interplay between trust, responsibility, and professional role identity.

Institutional trust is equally influential at a societal level. A U.S. survey examining public attitudes toward AI governance finds that trust in AI companies correlates with reduced support for slowing or regulating AI development, suggesting that some individuals believe industry self-regulation is sufficient when companies are considered trustworthy (Bullock et al., 2025). Conversely, individuals who trust government institutions more strongly tend to support stricter regulation. These findings show that public policy preferences around AI are not driven solely by risk assessments but also by trust in *the actors* responsible for managing those risks.

Taken together, these findings demonstrate that trust and trustworthiness in digital systems cannot be understood in technical terms alone. They are also shaped by cultural expectations, individual capabilities, professional identities, and institutional narratives. Engineering trustworthy systems remains essential -but achieving *societal* trust additionally requires transparent governance, education, and design approaches that account for human expectations and cognitive tendencies. Ultimately, trust is not simply a security property to be built; it is a socio-technical achievement that must be continually earned.

5. Use-Cases

Trust in digital systems and applications is, as mentioned, a multidimensional concept. It is therefore not sufficient to implement advanced cryptographic protocols and enforce zero-trust principles; trust must also be embedded in organizational processes, cultural norms, and human behavior. By describing the discussion using two application scenarios, the main objectives are: i) to illustrate theoretical concepts using concrete examples; ii) to highlight practical challenges that arise when technology encounters sociocultural diversity; and iii) to provide guidelines for designing trust frameworks that are both technically robust and socially resilient.

5.1 Cross-Border Energy Grid Interconnection

Energy networks are among the critical infrastructures, increasingly spanning multiple jurisdictions. From a technological point of view, securing these systems requires, among other things, encrypted SCADA communication, strong cryptographic protocols, Trusted Execution Environments (to protect calculations), and the enforcement of zero trust to prevent unauthorized access. These measures thus create a resilient technical foundation.

However, socio-cultural realities complicate this situation. Operators in different countries often have different training standards, operating cultures, and risk perceptions. Older teams may furthermore resist new security paradigms due to their long-standing dependence on traditional systems, while language barriers and differing regulations hinder collaboration. These factors can undermine even the most advanced technical controls.

5.2 Federated Healthcare Data Exchange

An additional example is the healthcare system. These systems are increasingly dependent on secure data exchange for diagnosis and research. From a technological perspective, this includes federated identity systems for cross-institutional authentication, cryptographic guarantees for data confidentiality, and TEEs to ensure data protection in AI-supported analyses. These mechanisms also form the backbone of trustworthiness here.

However, socio-cultural factors also increase complexity in this area. Medical professionals differ in terms of digital literacy, data privacy awareness, and compliance culture. Hospitals in rural areas may have fewer resources than urban facilities, leading to inconsistent security measures. Ethical standards and patient expectations regarding data privacy vary by region and influence the acceptance of new technologies.

5.3 Results and Discussion

The analysis of the two use cases illustrates how technological mechanisms and socioeconomic interventions combine to shape trust and trustworthiness in complex, interdependent infrastructures. In both areas, technological controls create verifiable, auditable evidence of trustworthiness, while socioeconomic measures determine whether this evidence is understood, accepted, and integrated into daily practice.

From a technological perspective, both use cases demonstrate that cryptographic security measures, identity systems, zero-trust architectures, and TEEs significantly strengthen the objective security posture of critical systems. In the energy sector, encrypted SCADA communications, robust key management processes, and attested workloads reduce the attack surface and improve data provenance. The enforcement of zero trust replaces implicit trust boundaries with dynamic, request-based authorizations, thereby limiting opportunities for lateral movement. These mechanisms lead to measurable improvements in cryptographic hygiene, interoperability, and forensic reconstructability. Similarly, in healthcare, federated identity systems and context-aware authorizations provide auditable access paths suitable for clinical audits, while TEEs protect sensitive data sets used in AI-powered diagnostic workflows. The result is an interoperable trust structure characterized by non-repudiation, minimized over privileges, and privacy-compliant analytics. In both sectors, the technical layer thus serves as a producer of verifiable evidence -certification proofs, signed telemetry, encryption suite inventories- that underpins institutional trustworthiness.

However, the results also make it clear that technological evidence alone is not sufficient to maintain trust in practice. Socioeconomic factors determine how stakeholders interpret technical controls, whether they are willing to follow new procedures, and how they assess the fairness and legitimacy of system behavior. In the energy use case, intercultural training harmonizes risk perception and reduces misinterpretations of alerts across languages and operating cultures. Standardized frameworks such as IEC 62443 and NIS2 create common expectations across national jurisdictions, while shared governance structures reduce ambiguity about roles and escalation paths. These measures are accompanied by measurable improvements, including fewer cross-border handover errors and a greater willingness to share indicators of compromise. In healthcare, initiatives to promote digital literacy improve the correct use of identity and security tools, reducing misrouting and misuse of emergency measures. Contextual guidelines enable rural hospitals to meet minimum legal requirements despite limited resources, preventing the emergence of unfair, two-tiered trust systems. Transparency materials for patients further improve perceived fairness and increase trust in data sharing and AI-assisted care. These socioeconomic measures thus translate technical trustworthiness into genuine social trust by promoting understanding, incentives, and accountability.

The combined results underscore a key finding: trust in critical digital infrastructures is created by aligning technical safeguards with sociocultural and organizational practices. Technology can determine what is secure, verifiable, and enforceable, but socioeconomic design determines whether these capabilities are accepted, understood, and defended. Sustainable trust therefore arises not solely from technical sophistication, but from the joint development of technical evidence, institutional governance, and human-centered engagement strategies.

Table 1: Trust & Trustworthy Aspects for Energy Grid and Federated Healthcare Use-Cases

Aspect	Cross-Border Energy Grid Interconnection	Federated Healthcare Data Exchange
Primary Goal	Secure and resilient energy distribution across national borders	Secure sharing of sensitive patient data across institutions
Technological Foundations	SCADA security and IoT integration; Cryptographic protocols; Trusted Execution Environments to control system integrity; Zero Trust for continuous verification:	Federated Identity Systems; Cryptographic guarantees for data confidentiality; Trusted Execution Environments for confidential AI-driven diagnostics;
Socio-Cultural Factors	Different training standards across countries; Language barriers;	Variability in digital literacy among medical staff; Cultural norms around patient privacy and consent; Resource disparities between rural and urban hospitals;
Trust Gaps and Challenges	Cascading failure affecting millions; Geopolitical tensions; Dependence on alignment of multiple national authorities; Limited interoperability in legacy SCADA systems; Potential misunderstanding of automated risk scores	Balancing innovation with privacy; Ensuring interoperability without compromising security; Persistent variation in resource availability; Uneven adoption of federated identity tools; Patient skepticism toward algorithmic decision-making;
Trust-Building Mechanisms	Alignment of cryptographic assurances with standardized cross-cultural operational practices	Alignment of technical privacy protections with clinician competence, hospital resources, and patient expectations

6. Conclusion and Future Work

In an increasingly interconnected and digitized world, trust and trustworthiness have become essential prerequisites for the resilience of critical infrastructures and the security of digital interactions. Yet despite significant advances in cryptographic safeguards, identity management, and zero-trust architectures, a persistent gap remains between technically engineered trustworthiness and the socio-cultural factors that shape how individuals and institutions actually perceive, interpret, and implement trust. Therefore, this work aims to bridge this gap by examining how measurable technical assurances interact with human behavior, cultural norms, and organizational practices across heterogeneous environments.

The analysis shows that technologies such as multi factor authentication, federated identity systems, trusted execution environments, and adaptive zero-trust mechanisms provide strong, verifiable evidence of trustworthiness by reducing attack surfaces, strengthening authentication, and enabling attestation-based assurance. However, these mechanisms only unfold their full effectiveness when complemented by socio-economic measures—including intercultural training, governance alignment, digital literacy programs, and transparent communication strategies—that enable stakeholders to understand, accept, and correctly apply technical controls.

Across all examined use cases, it becomes clear that sustainable trust is not a purely technical outcome but a socio-technical achievement. It emerges only when technical safeguards and human factors are consciously coordinated to reinforce one another, ensuring that engineered trustworthiness is matched by informed and confident user trust.

Acknowledgment

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 16KIS2239K SUSTAINET_guarDian). The authors alone are responsible for the content of the paper.

Ethics and Data Use Statement: This research used publicly available, anonymized secondary data. No identifiable personal information was accessed. Ethical approval was not required in accordance with institutional policies.

AI Assistance Statement: The authors used generative AI tools solely for linguistic editing and improving readability (e.g., grammar, phrasing, and structure). No AI system contributed to the formulation of the research questions, methodology, data interpretation, theoretical reasoning, or conclusions. All content was critically reviewed and approved by the authors, who take full responsibility for the manuscript.

References

Ackerhans, S., Wehkamp, K., Petzina, R., Dumitrescu, D., and Schulz, C., "Perceived Trust and Professional Identity Threat in AI-Based Clinical Decision Support Systems: Scenario-Based Experimental Study on AI Process Design Features", JMIR Publlications – Advancing Digital Health & Open Science, vol. 9, 2025. DOI: 10.2196/64266

Baron, N., Kidd, C., McNealy, J., Rainie, L., "Teens, Trust, and Technology in the Age of AI: Navigating Trust in Online Content", *Common Sense Media*, [online], Available at: https://www.commonsensemedia.org/sites/default/files/research/report/teens-trust-and-technology-in-the-age-of-ai_v2_web.pdf, 2025

Boston Institute of Analytics, "The Biggest Cyber Attacks of 2025: Lessons Learned and the Need for Cybersecurity Experts", [online], Available at: https://bostoninstituteofanalytics.org/blog/the-biggest-cyber-attacks-of-2025-lessons-learned-and-the-need-for-cybersecurity-experts/, [Accessed 2026-01-12]

Bullock, J.B., Pauketat, J.V.T., Huang, H., Wang, Y.-F-, and Anthis, J.R., "Public Opinion and the Rise of Digital Minds: Perceived Risk, Trust, and Regulation Support", *Public Performace & Management Rview, Generative AI and Public Administration: Opportunities and Challenges,* vol. 48, no. 6, 2025, DOI: 10.1080/15309576.2025.2495094

Cho, J.H., Chan, K., and Adali, S., "A Survey on Trust Modeling", *ACM Computing Surveys (CSUR),* vol. 48, no. 2, pp. 1—40, 2015, DOI: 10.1145/281559

Chokkanathan, K., Karpagavalli, S.M., Priyanka, G., Vanitha, K., Anitha, K., and Shenbagavalli, P., "AI-Driven Zero Trust Architecture: Enhancing Cyber-Security Resilience", *8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)*, Bengaluru, India, 2024, DOI: 10.1109/CSITSS64042.2024.10816746

Directive (EU) 2022/2555 of the European Parliament and of the Council of 12 December 2022 on measures for a for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive), Availbake at: https://eur-lex.europa.eu/eli/dir/2022/2555

Fettweis, G.P., Grünberg, P., Hentschel, T., and Köpsell, S., „Conceptualizing Trustworthiness and Trust in Communications ", *IEEE Communications Magazine*, vol. 63, no. 12, 2025, 10.1109/MCOM.001.2400383.

Finney, G., and Kindervag, J., "The Identity Cornerstone", in Project Zero Trust: A Story about a Strategy for Aligning Security and the Business, Wiley Data and Cybersecurity, pp. 57 – 72, 2023, DOI: 10.1002/9781394255146

Gambo, M.L., and Almulhen, A., "Zero Trust Architecture: A Systematic Literature Review", Journal of Network and System Management, vol. 34., no. 25, pp. 772 – 776, DOI: 10.1109/IAECST57965.2022.10062213

Gerlich, M., "Exploring Motivators for Trust in the Dichotomy of Human—AI Trust Dynamics", *social sciences*, vol. 13, no. 5, 2024, DOI: 10.3390/socsci13050251

Glazer, S., and Karpati, S., „The Role of Culture in Decision Making", *The Cutter Edge IT Journal,* vol. 27, no. 9, pp. 23—29, 2014. Lipps, C., Herbst, J., and Schotten, H.D. "How to Dance your Passwords: A Biometric MFA-scheme for Identification and Authentication of Individuals in IIoT Environments", *16th International Conference on Cyber Warfare and Security (ICCWS),* Cookeville, Tennessee, US, 2021.

Huynh, M.-T., and Aichner, T., „In generative artificial intelligence we trust: unpacking determinants and outcomes for cognitive trust ", *AI & Society,* vol. 40, pp. 5849—5869, 2025, DOI: 10.1007/s00146-025-02378-8

Lipps, C., Weinand, A., Krummacker, D., Fischer, C., and Schotten, H.D., „Proof of Concept for IoT Device Authentication Based on SRAM PUFs Using ATMEGA 2560-MCU", *1st International Conference on Data Intelligence and Security (ICDIS),* South Padre Island, TX, USA, 2018, DOI: 10.1109/ICDIS.2018.00013.

Lipps, C., Mallikarjun, S.B., Strufe, M., Heinz, C., Grimm, C., and Schotten, H.D., „Keep Private Networks Private: Secure Channel-PUFs, and Physical Layer Security by Linear Regression Enhanced Channel Profiles", *3rd International Conference on Data Intelligence and Security (ICDIS)*, Shenzhen, China, 2020, DOI: 10.1109/ICDIS50059.2020.00019.

Lipps, C., Baradie, S., Herbst, J., Armistead, L., and Schotten, H.D., "Cybersecurity in Industrial Automation and Control Systems: The Recent Attack of the Colonial Pipeline", in *Modelling Nation-state Information Warfare and Cyber-operations*, ACPIL, 2022, ISBN: ISBN: 978-1914587382

Luhmann, N., Davis, H. (translated by), Raffan, J. (translated by), Rooney, K. (translated by), King, M. (Editor), Morgner, C. (Editor), "Trust and Power", *Polity*, 2017, ISBN: 978-1509519453

Naprys, E., "Hackers pose as job seekers: opening a resume leads to ransomware" cybernews, [online], Available at: https://cybernews.com/security/hackers-target-recruiters-by-sending-malicious-resumes/, [Accessed 2026-01-15]

Pritchard, S., "The Evolving Cybersecurity Challenge for Critical Infrastructure", *Infosecurity Magazine,* [Online], Available at: https://www.infosecurity-magazine.com/news-features/cybersecurity-for-critical/, [Accessed 2026-01-12]

Pujari, S.R., and Hussain, M.A., "Human Factors in Cybersecurity: Behavioral Insights into Phishing and Social Engineering Attacks", *Nanotechnology Perceptions*, vol. 20, no. S15, ISSN: 1660-6795.

Regulation (EU) 2024/2847 of the European Parliament and of the Council of 23 October 2024 on horizontal cybersecurity requirements for products with digital elements and amending Regulations (EU) No 168/2013 and (EU) 2019/1020 and Directive (EU) 2020/1828 (Cyber Resilience Act)

Schneier, B., "Applied Cryptography: Protocols, Algorithms, and Source Code in C", Wiley & Sons, ISBN: 978-0471117094, 1995.

Shepherd, C., and Markantonakis, K., "Trusted Excecution Environments", Springer, 2024, ISBN: 978-3031555602

Sithungu, S., and Lipps, C., "Critical Infrastructure Security and the Role of AI: An Overview", 24th European Conference on Cyber Warfare and Security (ECCWS), Kaiserslautern, Germany, 2025, DOI: 0.34190/eccws.24.1.3770TrustTalk, "Trusting, Trustworthiness and Trust", [online], Available at: https://trusttalk.co/trusting-trustworthiness-and-trust/, [Accessed 2026-01-12],

Yates, J.F., and de Oliveira, S., "Culture and decision making", *Organizational Behaivor and Human Decision Process,* pp. 106—118, 2016, 10.1016/j.obhdp.2016.05.003.

Yamamoto, Y., "A morality based on trust: Some reflections on japanese morality", *Philosophy East and West*, vol. 40, no. 4, pp. 451—469, 1990, DOI: 10.2307/1399351

Van Der Ham, J., "Toward a Better Understanding of "Cybersecurity", Digital Threats: Research and Practice, vol. 2, no. 3, pp 1-3, DOI: 10.1145/3442445

Q-Learning Model for Proportionality Assessment in Military Operations

Clara Maathuis
Open University of the Netherlands, Heerlen, The Netherlands

clara.maathuis@ou.nl

Abstract: In the context of military operations, accurate and transparent proportionality assessment is essential to ensure compliance with international humanitarian law. On this behalf, this research presents and evaluates two Q-learning models designed to build the proportionality assessment in military operations. In this sense, the first model considers collateral damage exclusively in physical terms (excluding psychological harm), while the second model explicitly integrates psychological damage as part of the collateral damage effects. Both models encode operational rules as multi-attribute states encompassing injury severity, fatalities, object damage, and military advantage, differing only in the inclusion of psychological factors. From training and simulation results, it can be seen that this approach provides a valuable classification approach for proportional and disproportional outcomes within their respective scenario sets. This shows that AI (Artificial Intelligence) provides effective methods and techniques that allow both modelling and the expansion of existing definitions and perspectives of existing challenging concepts in the uncertain and dynamic space of the military domains while accounting and respecting legal and ethical considerations in order to build responsible and trustworthy military AI systems.

Keywords: Military operations, Military targeting, Proportionality, Artificial intelligence, Reinforcement learning, Q-learning

1. Introduction

Motto: "The fragility of freedom is the simplest and deepest lesson of my life and work ." (Albert Einstein)

Recent years witnessed unprecedented advancements in the Artificial Intelligence (AI) domain, driven by breakthroughs in machine learning, deep neural networks, and reinforcement learning paradigms. These applications have transcended traditional data analysis tasks to enable autonomous perception, reasoning, and decision-making across a variety of complex domains (Sarker, 2021; Hussain, Rahman & Ali, 2024). In particular, the increased applicability of AI-enabled systems, ranging from unmanned aerial vehicles to automated sensor fusion platforms, has catalysed their integration into military operations (Criollo, Mena-Arciniega & Xing, 2024; Singh, 2024). At the same time, worldwide, various military organizations are exploring how AI can augment situational awareness, optimize logistical workflows, and support command and control (C2) architectures, bringing a new era in which intelligent agents play an increasingly central role in strategic and tactical decision loops (Lingel et al., 2020; Tóth & Farkas, 2023; Toroi, 2024). Within the military operational context, numerous decision-making processes stand to benefit from AI-driven automation. This points out to solutions such as the ones developed and deployed for target identification and selection, Rules of Engagement adjudication, and fire-control authorization represent high-stakes tasks that demand rapid, consistent judgments under uncertain and time critical conditions (Jahn, 2019; Schmitt, M. N., & Schauss, 2019; Lekea, Lekeas & Topalnakos, 2023). At the same time, machine learning-based classifiers assist in discriminating combatants from non-combatants and in prioritizing threats based on mission objectives, while automated planning algorithms can generate and evaluate courses of action in parallel, reducing human cognitive burden (Ahmed, 2022; Montasari, 2024; Ray, 2024). By embedding AI into these activities, armed forces aim to enhance responsiveness, reduce unintended effects such as collateral damage, and free human operators to focus on higher-order strategic reasoning.

A particularly sensitive application of AI in warfare lies in autonomous targeting systems, where adherence to International Humanitarian Law (IHL) is an important requirement (Sassoli, 2014; Khali & Raj, 2024). Central to IHL is the principle of proportionality, codified in Additional Protocol I to the Geneva Conventions, most notably Article 51(5)(b), which prohibits attacks "which may be expected to cause incidental loss of civilian life, injury to civilians, or damage to civilian objects … which would be excessive in relation to the concrete and direct military advantage anticipated," and Article 57's requirement to take "all feasible precautions" to minimize harm to civilians (Benvenisti, 2022; Al-Fatlawi, Tamimi & Al-Tamimi, 2024). Responsible AI solutions for targeting should therefore internalize proportionality assessments, ensuring that algorithmic decisions respect these legal norms. Hence, designing AI agents that can systematically evaluate collateral damage against anticipated military advantage is essential to upholding lawful conduct in future conflict scenarios. To this end, integrating legal and ethical perspectives is indispensable when engineering AI systems for

proportionality assessment and targeting support, as these systems operate at the intersection of life-and-death decisions and binding international norms. In addition to the legal perspective, ethically, responsible AI frameworks compel respect for human dignity, civilian protection, and the minimization of unintended harm, necessitating transparent value-alignment mechanisms, human-in-the-loop safeguards, and accountability provisions. In this context, despite extensive theoretical and conceptual research on collateral damage and proportionality assessment spanning legal treatises, ethical frameworks, philosophical analyses of jus in bello (Neuman, 2004; Henderson, I., & Reece, 2018; Fard & Maathuis, 2021; Cohen, A., & Zlotogorski, 2021; Maathuis & Chockalingam, 2023), and technical proposals for decision models, a gap remains in translating these insights into intelligent and adaptive responsible AI systems .

Traditional studies have systematically catalogued normative criteria, harm quantification methodologies, and decision heuristics, yet few have bridged the divide to operationalize proportionality as a dynamic, learnable function within reinforcement-learning or hybrid control architectures. As a result, existing AI efforts often rely on simplistic threshold-based filters, lacking the capacity to generalize across evolving tactical contexts or to calibrate judgments in the face of novel collateral damage profiles. Addressing this gap calls for transdisciplinary research that merges rigorous legal-ethical formalization with AI techniques, thereby creating AI systems capable of real-time, context-aware proportionality assessments that learn and adapt while remaining legally and morally anchored.

To this end, this research aims to develop a Q-learning–based decision-support system (Watkins & Dayan, 1992) that encodes both legal norms and ethical imperatives into the proportionality assessment process for military targeting. By modelling proportionality as a Markov decision process (MDP) (Bellman, 1957), this approach represents each engagement scenario as a state vector of collateral damage attributes and military advantage, and uses tabular Q-learning to learn an optimal action-value function Q(s,a) mapping these states to Proportional or Disproportional decisions. Here, the agent's exploration-exploitation policy is governed by an ϵ-greedy strategy that ensures comprehensive sampling of legally relevant scenario permutations before converging on lawful and ethical targeting recommendations. On this behalf, to investigate the impact of different collateral damage conceptualizations, two experimental and simulation cases are defined. In Case 1, collateral damage encompasses only physical injury, death, and object destruction, thereby isolating the agent's ability to enforce legal proportionality under traditional kinetic effects. In Case 2, the state definition is extended to include psychological harm as well by evolving ethical recognition of non-lethal but debilitating effects on civilian populations, thereby increasing the rule complexity and ethical nuance the Q-learning agent must internalize (Maathuis, 2022). By comparing learning dynamics and policy formation across these two cases, this research aims to demonstrate how adaptive reinforcement learning architectures can be tuned to respect core legal thresholds while integrating emerging ethical dimensions of collateral damage for building a responsible AI model targeting decision-making support.

The remainder of this article is structured as follows. Section 2 introduces the background of this research and discusses related studies conducted in this domain. Section 3 presents the methodology followed to achieve the aim of this research. Section 4 presents the design of the model proposed. Section 5 presents the evaluation process and the evaluation results obtained. At the end, in Section 6 are discussed concluding remarks and a series of future research perspectives are provided.

2. Related Research

Q-learning is a fundamental algorithm within the RL (Reinforcement Learning) paradigm designed to enable agents to learn optimal decision-making policies through interaction with their environment (Watkins & Dayan, 1992; Clifton & Laber, 2020). In this paradigm, an agent sequentially observes the state of the environment, selects actions, and receives feedback in the form of rewards or penalties, with the goal of maximizing cumulative future rewards. Specifically, Q-learning is distinguished by being a value-based, model-free, and off-policy algorithm, which implies that it does not require knowledge of the environment's dynamics and is not depending on the policy currently being followed. Instead, it directly estimates the optimal action-value function, or Q-function, which quantifies the expected cumulative reward for taking a particular action in a given state and subsequently following the best possible strategy (Tan, Yan & Guan, 2017). Through iterative updates that use the Bellman equation, the agent refines these Q-values, gradually converging towards the optimal policy that maximizes long-term returns.

The main components of the Q-learning algorithm include the agent, the state space representing all possible situations the agent might encounter, and the action set from which the agent chooses behaviours. After every action taken, the agent receives a scalar reward reflecting the immediate benefit or cost of that decision,

which is used to update the Q-values stored in a tabular or functional form that maps state-action pairs to expected returns. Accordingly, the core parameters relevant to this learning process are as follows: the learning rate which controls the influence of new experiences relative to existing knowledge, while the discount factor balances the importance of immediate against future rewards (Oliehoek, Spaan & Vlassis, 2008; Sadhu & Konar, 2020). Moreover, an exploration-exploitation strategy, typically epsilon-greedy, manages the trade-off between trying unfamiliar actions to discover potentially better rewards and exploiting current knowledge to maximize returns. In addition, the model-free and off-policy nature that the Q-learning algorithm has contributes to both its versatility and its robustness across various societal applications such as autonomous robotics, game playing, and decision support systems (Zhang & Mo, 2021; Rajasekhar, Radhakrishnan & Samsudeen, 2025). Despite these advantages, classical Q-learning struggles with scalability when dealing with large or continuous state spaces since tabular representations become infeasible. This limitation has motivated the development of deep Q-learning, which employs the power of neural networks as function approximators to generalize Q-values beyond discrete states, thereby enabling efficient learning in complex environments (Arulkumaran et al., 2017; Jang et al., 2019).

One of the most established applications is in active protection systems (APS) for armoured vehicles, where Q-learning is utilized to optimize the deployment of soft-kill countermeasures against guided high-mobility threats, such as anti-tank missiles (Rajagopalan, 2018). This real-time decision-making capability enhances survivability by adaptively responding to diverse attack vectors. Beyond physical defence, Q-learning is implemented in cyber defence systems to autonomously detect, analyse, and mitigate complex cyber threats such as advanced persistent threats (APTs) and zero-day exploits within military networks (Stellios, Kotzanikolaou & Psarakis, 2019; Hernández-Rivas, Morales-Rocha & Sánchez-Solís, 2024). Other applications include autonomous vehicle control (e.g., drones, tanks, unmanned surface vessels) where Q-learning guides navigation, obstacle avoidance, and mission task execution without explicit programming, thereby allowing adaptive responses to evolving battlefield conditions (Zhang et al., 2020; Chen et al., 2024).

In military decision support and C2, Q-learning plays is often use in optimizing weapon selection, threat evaluation, and resource allocation. This enable systems to learn effective strategies for weapon assignment, coordinating multi-armed platforms to prioritize and engage threats efficiently, which is critical in fast-paced combat scenarios (Cheng, Chen & Gong, 2021; Wang, Zhang & Wang, 2024). For instance, X applies it in battlefield simulations to train agents for tactical decision-making that balances mission objectives with minimizing casualties (Costa et al., 2025; Park & Shim, 2025). Furthermore, Q-learning algorithms assist in logistics planning by optimizing the distribution of supplies, personnel, and equipment across operational theatres to ensure mission readiness and sustainability (Singh, Mandal & Nayak, 2024). At the same time, various reinforcement learning techniques have also been used in air combat simulation and pilot training, where Q-learning networks integrated with deep learning help build intelligent virtual pilots capable of self-learning tactics and confronting human pilots in simulated engagements (Ruan, Duan & Deng, 2022).

Other applications include applications for intelligent reconnaissance and surveillance systems, where agents autonomously explore and gather intelligence while adapting to changing environments; multi-agent cooperation for coordinated actions in complex missions; electronic warfare strategies for adaptive signal jamming and countermeasures; autonomous underwater vehicle task scheduling; adaptive encryption and cyberthreat mitigation in sensor networks; and battlefield robotics for target acquisition and engagement. These applications show the versatility of Q-learning in enabling autonomous, adaptive, and optimized decision-making across a broad range of defence systems, enhancing operational effectiveness and resilience in complex and unpredictable combat scenarios (Maathuis, Pieters & van den Berg, 2018; Abro et al., 2024; Premakumari et al., 2025). However, challenges remain in algorithm stability, interpretability of learned policies, adversarial robustness, and compliance with legal and ethical standards for autonomous military systems.

3. Research Methodology

To build a Q-learning model for proportionality assessment in military operations, this research adopts methodological principles of the Design Science Research (DSR) paradigm as articulated by (Kuechler & Vaishnavi, 2012; Peffers, Tuunanen & Niehaves, 2018) which emphasizes the iterative creation and rigorous evaluation of innovative artifacts to solve relevant, real-world problems. Following the DSR process framework, this research commences with the identification and motivation of the proportionality assessment challenges in military operations, highlighting the need for a decision-support mechanism capable of encoding complex legal and ethical humanitarian rules (Maathuis, Pieters & Van Den Berg, 2018b) which serves as the

base for the data used in this research. To this end, a Q-learning reinforcement learning approach is considered to build the proportionality assessment model that includes a broader perspective on the collateral damage component of this principle, i.e., the inclusion of psychological effects or harm. The design and development phase involved constructing this Q-learning agent with calibrated hyperparameters reflective of the deterministic rule complexity, followed by an in-depth performance evaluation through simulation on representative datasets. This ensures the artifact's functional robustness and its conforming to the intended decision criteria are fulfilling the DSR guidelines so that the artifact must demonstrates its effectiveness and utility within the military targeting decision-making process.

Furthermore, the model's evaluation incorporates metrics such as convergence dynamics, reward stability, and classification accuracy, augmented by confusion matrices for clear interpretability to support the transparency dimension of the responsible AI perspective governing this research (Maathuis & Chockalingam, 2022). On this behalf, the results are analysed in relation to learning-theoretic implications, linking hyperparameter configurations with convergence behaviour to inform methodological tuning. By iteratively refining the Q-learning model and documenting important modelling choices, this research contributes to the existing body of knowledge that aims at developing responsible and trustworthy military AI systems and existing practical efforts dedicated to building intelligent and adaptive operational AI systems.

4. Model

To build the Q-learning agent model that encodes proportionality as a decision-support system for military targeting, the assessment process is modelled as a Markov decision process since the outcomes are partly random and partly under the control of an agent. This framework consists of states, actions, transition probabilities between states after taking actions, rewards for transitions, and a discount factor for future rewards. The MDP assumes the Markov property, meaning the next state depends only on the current state and action, not on the history. In this process, the agent observes a discrete state s reflecting collateral-damage and military-advantage components, selects a binary action a∈{P for proportional and DP for disproportional} and receives a scalar reward r that guides off-policy updates of its action-value function Q(s,a)Q(s,a). Then, the learning proceeds using the Bellman equation:

$$Q(s,a) \leftarrow (1-\alpha)\,Q(s,a) + \alpha\big(r + \gamma \max_{a'} Q(s',a')\big)$$

with a tabular Q-table initialized to zeros. An ϵ (epsilon)-greedy policy balances exploration and exploitation, ensuring the agent samples all scenario–action pairs while progressively favouring optimal classifications. Further, taking into account the two modelling perspectives where psychological damage is not considered (i.e., Case 1) and considered (i.e., Case 2) as being part of the collateral damage component, both modelling cases are further discussed.

In Case 1, where psychological harm is excluded from collateral damage, the state is the 4-tuple (CD_injury_no_psycho, CD_death, CD_object, MA) meaning collateral effects that do not imply psychological harm, the ones that imply civilian loss of life, the ones that imply damage or destruction to civilian objects, and Military Advantage, where each of the first two dimensions takes values {L, M, H} meaning Low, Medium, and High, the third {Y,N} which means Yes and No, and the fourth {L, M, H}. This implies 3×3×2×3=54 unique states based on the inclusion or exclusion of psychological harm. At each training step, the agent samples a state uniformly, selects P or DP using ϵ-greedy (ϵ=0.1), and receives a reward of +10 if the choice matches the expert rule and –10 otherwise. In this process, the updates apply with learning rate α=0.1 and discount factor γ=0.9 iterated over several thousand episodes to populate the Q-table with the expected returns for each state–action pair.

In Case 2, psychological injury is incorporated in the collateral damage component by replacing the first feature with CD_injury_with_psycho, thereby enriching the state semantics while keeping the action set unchanged . The added complexity demands hyperparameter recalibration: α is reduced to 0.01 and ϵ is increased to 0.2 to slow Q-value uptake and prolong exploration across the expanded rule boundary . The learning loop and Bellman updates remain identical, but these conservative settings ensure that the agent visits each of the 54 states sufficiently often and refines its Q-table smoothly, enabling reliable encoding of both physical and psychological collateral-damage assessments in a single, unified tabular model.

5. Evaluation

In the first case, agent was trained on a fully deterministic rule set mapping collateral-damage and military-advantage features to a binary proportionality decision (P vs. DP). Using a learning rate of $\alpha = 0.1$, discount factor $\gamma = 0.9$, and exploration rate $\varepsilon = 0.1$ over 1 000 episodes, the total reward curve exhibits a rapid ascent from initially negative values corresponding to random exploratory actions to a stable plateau around +45 to +55 per episode by ~100 episodes. This indicates that the agent consistently selects correct actions thereafter. The resulting Q-values yield perfect classification (accuracy = 100%), as confirmed by a confusion matrix with 20 DP and 34 P cases all correctly identified (Figure 1). The smooth convergence of the reward trajectory and the clear separation of Q(P) vs. Q(DP) in sampled states demonstrate that the chosen hyperparameters strike an effective balance between exploration and exploitation, enabling the agent to fully internalize the rule structure with minimal oscillation once learned (Figure 2).

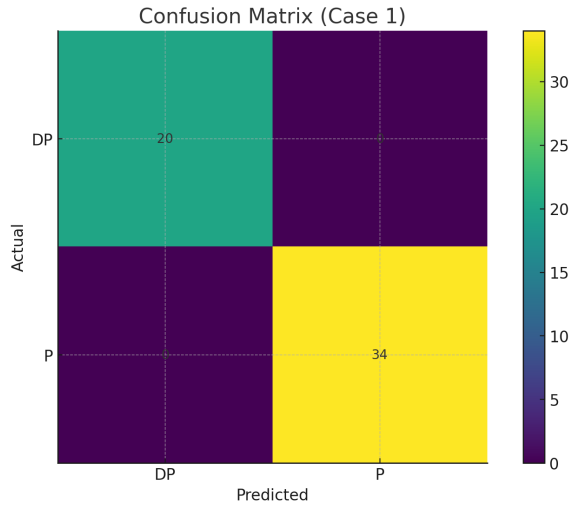


Figure 1: Confusion matrix for Case 1

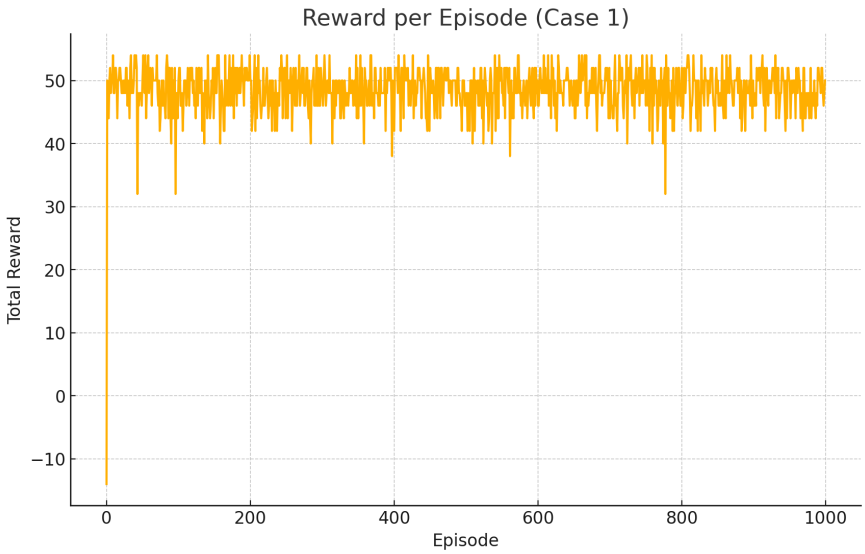


Figure 2: Reward per episode for Case 1

In the second case, when extending the rule set to include psychosocial injury, a targeted hyperparameter sweep over learning rates ($\alpha \in \{0.01, 0.05, 0.1\}$) and exploration rates ($\varepsilon \in \{0.01, 0.1, 0.2\}$) is executed. The optimal configuration ($\alpha = 0.01$, $\varepsilon = 0.2$, $\gamma = 0.9$) achieved consistent 100% accuracy while yielding a reward curve that, although more variable early on, stabilized by ~200 episodes in the +40 to +50 range. The lower learning rate reduced Q-value update magnitudes, smoothing convergence and mitigating large reward swings, while a moderate exploration rate ensured sufficient sampling of state-action pairs. The confusion matrix (22 DP and 32 P correctly classified) shows that, even with additional feature complexity, careful tuning

of α and ε delivers robust learning (Figure 3). Overall, these results confirm that Q-learning can reliably encode deterministic proportionality rules and that hyperparameter selection critically influences the speed and stability of convergence (Figure 4).

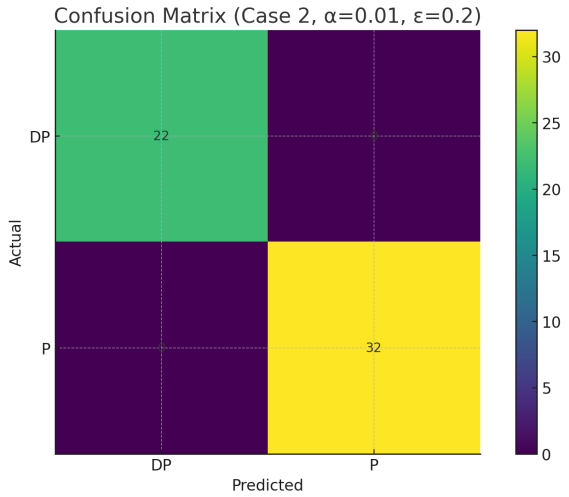


Figure 3: Confusion matrix for Case 2

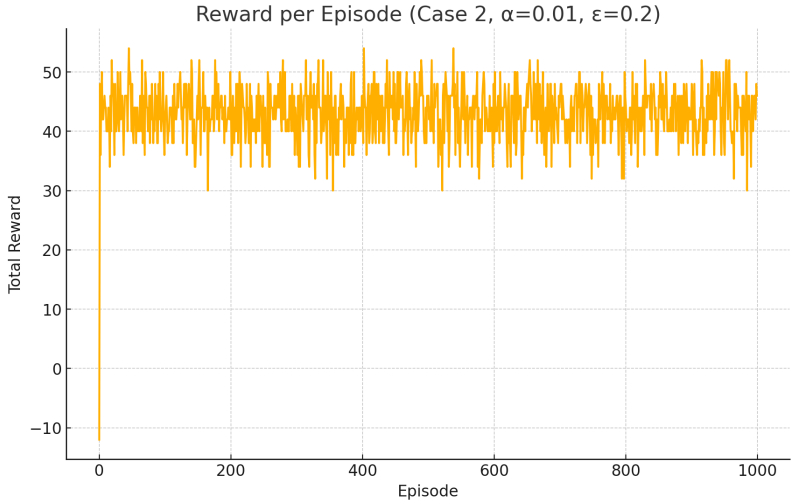


Figure 4: Reward per episode for Case 2

Although both cases achieved a good classification, their learning dynamics exhibit distinct trade-offs driven by problem complexity and hyperparameterization. Specifically, in Case 1, a relatively high learning rate (α = 0.1) and moderate exploration (ε = 0.1) enabled the agent to rapidly accrue positive reward, stabilizing by roughly 100 episodes, and converging on the correct Q-values with minimal oscillation. At the same time, the enriched feature space of Case 2 required a much lower learning rate (α = 0.01) paired with increased exploration (ε = 0.2) to smoothly integrate the additional rule patterns: although rewards initially fluctuated more broadly and convergence took closer to 200 episodes, the softer updates and sustained exploration prevented premature convergence to suboptimal actions. Hence, while in both cases accuracy is high, Case 1 had a speed of acquisition under simpler rule sets, whereas Case 2 reflected the stabilizing role of conservative learning rates and thorough exploration when encoding more complex decision boundaries.

From a learning-theoretical standpoint, both Case 1 and Case 2 demonstrate Q-learning's capacity to internalize deterministic proportionality mappings, but their Q-value trajectories and update semantics differ markedly. In Case 1, a higher learning rate (α = 0.1) yields rapid Q-value growth for the correct actions, reflected by a steep reward curve and large separation between Q(P) and Q(DP) after only ≈100 episodes, whereas in Case 2 the reduced α (0.01) produces more gradual, finely graded increments to Q-values. This conservative updating, combined with elevated exploration (ε = 0.2), preserves stochastic sampling longer into training, preventing overly confident Q-values before the full rule set is observed and smoothing the reward

signal until ≈200 episodes. In practical terms this means that one should calibrate α to the complexity of the decision space: simpler rule sets tolerate, and benefit from, larger step-sizes, while richer feature combinations favour smaller α to avoid oscillatory Q-updates. Moreover, monitoring the gap between the highest and second-highest Q-values per state can serve as a convergence diagnostic: narrow margins suggest the need for further exploration or slower learning.

6. Conclusions

This research introduces a Q-learning modelling approach to conduct the proportionality assessment in military operations considering various levels of problem complexity. In the first case scenario without psychosocial injuries, a relatively high learning rate and moderate exploration enabled rapid convergence and stable, perfect classification of proportionality decisions, underscoring the agent's ability to quickly learn deterministic rules under constrained feature spaces. When the feature set was extended to include psychosocial injury considerations, careful hyperparameter tuning toward a lower learning rate and increased exploration was necessary to achieve smooth convergence and maintain perfect accuracy. This illustrates the delicate balance required between learning speed and stability, particularly as the decision boundary complexity increases. The differentiated dynamics of Q-value updates observed between the two cases provides practical insights for tailoring reinforcement learning strategies to the complexity of military proportionality constraints, ensuring reliable encoding of operational, legal, and ethical considerations in autonomous military decision-making support systems.

From a broader AI and military operational perspective, these findings support the viability of reinforcement learning as a tool for encoding and modelling the proportionality assessment principle in settings that imply both a high level of autonomy or human-in-the-loop targeting decisions (Maathuis, 2024). The results obtained align with emerging research advocating for explainable and responsible AI systems capable of managing uncertainty and multidimensional rule sets in legally and ethically critical domains like military engagement. In addition, by enabling adaptive, robust classification through methodical hyperparameter calibration, Q-learning offers a promising pathway toward autonomous systems that respect legal and moral boundaries in warfare. From here, future implementations could consider integrating these models with function approximators such as deep Q-networks to generalize beyond deterministic rules and embed even more real-world uncertainty elements and partial observability. Specifically, two future research perspectives emerge from this research. First, investigating the resilience of such models under conditions of uncertainty and partial observability typical of combat environments, possibly through introducing controlled noise or incomplete information during training. And second, and extending the approach toward function approximation methods, such as deep reinforcement learning, to enable scalable generalization and real-time adaptability in complex, dynamic operational theatres beyond rigid constraints. These directions aim to enhance the practical utility as well as legal and ethical reliability of AI-driven proportionality assessments in future military decision support systems.

Declaration: For this research, no ethical clearance is required and no AI tools were used in the creation of this article.

References

Abro, G. E. M., Ali, Z. A., & Masood, R. J. (2024). Synergistic UAV motion: A comprehensive review on advancing multi-agent coordination. *ICCK Transactions on Sensing, Communication, and Control*, *1*(2), 72-88.

Ahmed, N. U. (2022). Integrating machine learning in military intelligence process: study of futuristic approaches towards human-machine collaboration. *NDC e-journal*, *2*(1), 59-89.

Al-Fatlawi, A. A., Tamimi, K., & Al-Tamimi, M. J. A. (2024). Feasible Precautions: A Legal Study in the Stable and Variable Concept Under International Humanitarian Law. *Journal of college of Law for Legal and Political Sciences*, *10*(37), 598-629.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE signal processing magazine*, *34*(6), 26-38.

Bellman, R. (1957). A Markovian decision process. *Journal of mathematics and mechanics*, 679-684.

Benvenisti, E. (2022). The Birth and Life of the Definition of Military Objectives. *International & Comparative Law Quarterly*, *71*(2), 269-295.

Chen, Y., Ji, C., Cai, Y., Yan, T., & Su, B. (2024). Deep reinforcement learning in autonomous car path planning and control: A survey. *arXiv preprint arXiv:2404.00340*.

Cheng, Q., Chen, D., & Gong, J. (2021). Weapon-target assignment of ballistic missiles based on Q-learning and genetic algorithm. In *2021 IEEE International Conference on Unmanned Systems (ICUS)* (pp. 908-912). IEEE.

Cohen., D. (2021). *Proportionality in international humanitarian law: Consequences, precautions, and procedures*. Oxford University Press.

Clifton, J., & Laber, E. (2020). Q-learning: Theory and applications. *Annual Review of Statistics and Its Application*, *7*(1), 279-301.

Costa, A. N., Dantas, J. P., Scukins, E., Medeiros, F. L., & Ã–gren, P. (2025). Simulation and Machine Learning in Beyond Visual Range Air Combat: A Survey. *IEEE Access*.

Criollo, L., Mena-Arciniega, C., & Xing, S. (2024). Classification, military applications, and opportunities of unmanned aerial vehicles. *Aviation*, *28*(2), 115-127.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information systems*, *13*(6), 3.

Henderson, I., & Reece, K. (2018). Proportionality under international humanitarian law: the reasonable military commander standard and reverberating effects. *Vand. J. Transnat'l L.*, *51*, 835.

Hussain, M. D., Rahman, M. H., & Ali, N. M. (2024). Artificial intelligence and machine learning enhance robot decision-making adaptability and learning capabilities across various domains. *International Journal of Science and Engineering*, *1*(3), 14-27.

Hernández-Rivas, A., Morales-Rocha, V., & Sánchez-Solís, J. P. (2024). Towards autonomous cybersecurity: A comparative analysis of agnostic and hybrid AI approaches for advanced persistent threat detection. In *Innovative Applications of Artificial Neural Networks to Data Analytics and Signal Processing* (pp. 181-219). Cham: Springer Nature Switzerland.

Jahn, J. L. (2019). Shifting the safety rules paradigm: Introducing doctrine to US wildland firefighting operations. *Safety science*, *115*, 237-246.

Jang, B., Kim, M., Harerimana, G., & Kim, J. W. (2019). Q-learning algorithms: A comprehensive classification and applications. *IEEE access*, *7*, 133653-133667.

Khalil, A., & Raj, S. A. K. (2024). Deployment of autonomous weapon systems in the warfare: Addressing accountability gaps and reformulating international criminal law. *Balkan Social Science Review*, *23*(23), 261-285.

Lekea, I., Lekeas, G., & Topalnakos, P. (2023). Exploring Enhanced Military Ethics and Legal Compliance through Automated Insights: An Experiment on Military Decision-making in Extremis. *Conatus-Journal of Philosophy*, *8*(2), 345-372.

Lingel, S., Hagen, J., Hastings, E., Lee, M., Sargent, M., Walsh, M., ... & Blancett, D. (2020). Joint all domain command and control for modern warfare: an analytic framework for identifying and developing artificial intelligence applications.

Maathuis, C. (2022). An Outlook of Digital Twins in Offensive Military Cyber Operations. In *European Conference on the Impact of Artificial Intelligence and Robotics* (Vol. 4, No. 1, pp. 45-53).

Maathuis, C., & Chockalingam, S. (2022). Responsible digital security behaviour: Definition and assessment model. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1).

Maathuis, C., & Chockalingam, S. (2023). Modelling the influential factors embedded in the proportionality assessment in military operations. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 218-226).

Maathuis, C., Pieters, W., & van den Berg, J. (2018). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C., Pieters, W., & Van Den Berg, J. (2018b). A computational ontology for cyber operations. In *Proceedings of the 17th European Conference on Cyber Warfare and Security* (pp. 278-288).

Maathuis, C. (2024). Trustworthy Human-Autonomy Teaming for Proportionality Assessment in Military Operations. In *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)* (pp. 1-8). IEEE. Montasari, R. (2024). Addressing Ethical, Legal, Technical, and Operational Challenges in Counterterrorism with Machine Learning: Recommendations and Strategies. In *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses* (pp. 199-226). Cham: Springer International Publishing.

Neuman, N. (2004). Applying the rule of proportionality: force protection and cumulative assessment in international law and morality. *Yearbook of International Humanitarian Law*, *7*, 79-112.

Oliehoek, F. A., Spaan, M. T., & Vlassis, N. (2008). Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, *32*, 289-353.

Park, K., & Shim, S. (2025). Intelligent Counterforce Allocation Method Using Multi-Agent Reinforcement Learning for Ground Operations. *IEEE Access*.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Premakumari, S. B. N., Sundaram, G., Rivera, M., Wheeler, P., & Guzmán, R. E. P. (2025). Reinforcement Q-Learning-Based Adaptive Encryption Model for Cyberthreat Mitigation in Wireless Sensor Networks. *Sensors*, *25*(7), 2056.

Rajagopalan, A. (2018). Active protection system soft-kill using q-learning. In *International Conference on Science and Innovation for Land Power, Australia Defence Science and Technology*.

Rajasekhar, N., Radhakrishnan, T. K., & Samsudeen, N. (2025). Exploring reinforcement learning in process control: a comprehensive survey. *International Journal of Systems Science*, 1-30.

Ray, P. P. (2024). A Review of TRiSM Frameworks in Artificial Intelligence Systems: Fundamentals, Taxonomy, Use Cases, Key Challenges and Future Directions. *Authorea Preprints*.

Ruan, W., Duan, H., & Deng, Y. (2022). Autonomous maneuver decisions via transfer learning pigeon-inspired optimization for UCAVs in dogfight engagements. *IEEE/CAA Journal of Automatica Sinica*, *9*(9), 1639-1657.

Sadhu, A. K., & Konar, A. (2020). *Multi-agent coordination: A reinforcement learning approach*. John Wiley & Sons.

Sarker, I. H. (2021). Data science and analytics: an overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, *2*(5), 377.

Sassoli, M. (2014). Autonomous weapons and international humanitarian law: Advantages, open technical questions and legal issues to be clarified. *International Law Studies*, *90*(1), 1.

Schmitt, M. N., & Schauss, M. (2019). Uncertainty in the law of targeting: towards a cognitive framework. *Harv. Nat'l Sec. J.*, *10*, 148.

Singh, B. (2024). Unmanned aircraft systems (UAS), surveillance, risk management to cybersecurity and legal regulation landscape: unraveling the future analysis, challenges, demand, and benefits in the high sky exploring the strange new world. *Unmanned aircraft systems*, 313-354.

Singh, R., Mandal, P., & Nayak, S. (2024). Reinforcement Learning Approach in Supply Chain Management: A Review. *How Machine Learning is Innovating Today's World: A Concise Technical Guide*, 271-302.

Stellios, I., Kotzanikolaou, P., & Psarakis, M. (2019). Advanced persistent threats and zero-day exploits in industrial Internet of Things. In *Security and Privacy Trends in the Industrial Internet of Things* (pp. 47-68). Cham: Springer International Publishing.

Tan, F., Yan, P., & Guan, X. (2017). Deep reinforcement learning: From Q-learning to deep Q-learning. In *International Conference on Neural Information Processing* (pp. 475-483). Cham: Springer International Publishing.

Toroi, G. I. (2024). Rethinking military command and control systems. *Bulletin of" Carol I" National Defence University (EN)*, *4*(04), 88-112.

Tóth, A., & Farkas, T. (2023). Opportunities and Directions for the Evolution of Command and Control Systems in the Context of Multi-domain Operations.

Wang, X., Zhang, Y., & Wang, G. (2024). Target assignment for multiple stages of weapons systems using a deep Q-learning network and a modified artificial bee colony method. *Computers and Electrical Engineering*, *118*, 109378.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3), 279-292.

Zhang, L. A., Xu, J., Gold, D., Hagen, J., Kochhar, A. K., Lohn, A. J., & Osoba, O. A. (2020). *Air dominance through machine learning: A preliminary exploration of artificial intelligence-assisted mission planning* (No. RR4311RC).

Zhang, T., & Mo, H. (2021). Reinforcement learning for robot research: A comprehensive review and open issues. *International Journal of Advanced Robotic Systems*, *18*(3), 17298814211007305.

Proportionality Assessment in Military Operations Based on SARSA and PPO Models

Clara Maathuis
Open University of the Netherlands, Heerlen, The Netherlands

clara.maathuis@ou.nl

Abstract: This research proposes a novel decision-support framework for proportionality assessment in military operations integrating on-policy and policy-gradient reinforcement-learning methods to encode expert rules and automatically classify engagement scenarios. On this behalf, two modelling perspectives are considered implementing SARSA (State-Action-Reward-State-Action) and PPO (Proximal Policy Optimization) algorithms, and two approaches are adopted, i.e., in order to consider or not the integration of psychological effects or harm as part of collateral damage. Further, various optimization methods and simulation scenarios are considered in order to understand the effectiveness and robustness of the modelling techniques developed. From the results obtained, it can be seen that while SARSA achieves rapid reward stabilization but exhibits limited accuracy due to potential bootstrapping bias and insufficient exploration in larger state spaces, PPO's clipped surrogate updates yield robust, monotonic improvement, consistently realizing high classification accuracy across both cases, albeit over longer training horizons. To this end, a comparative analysis is conducted based on simulation results, learning curves, Q-value/policy distributions, and confusion matrices to illustrate each algorithm's strengths and limitations. Hence, this research demonstrates the viability of reinforcement learning models as transparent, adaptable tools for proportionality assessment for supporting real-time operational decisions.

Keywords: Military operations, Proportionality, Artificial intelligence, Reinforcement learning, SARSA, PPO

1. Introduction

> ***Motto:*** *"The unleashed power of the atom has changed everything save our modes of thinking and we thus drift toward unparalleled catastrophe." (Albert Einstein)*

The rapid development of artificial intelligence (AI) technologies over the past decade has driven an unprecedented wave of innovation across all societal domains. Advances in machine learning, particularly deep neural networks and reinforcement-learning frameworks, enabled systems to perceive complex environments, infer latent patterns, and make autonomous decisions at speeds exceeding human capabilities (Matsuo et al., 2020; Taye, 2023; Chen et al., 2024; Maathuis, 2022; Govinda, Brik & Harous, 2025). In the military domain, these breakthroughs translated into a new generation of smart sensors, unmanned platforms, and decision-support systems that enhance situational awareness, optimize logistics, and automate routine Command and Control (C2) functions (Cole et al., 2024; Probasco et al., 2025). From autonomous reconnaissance drones that navigate contested airspace to sensor-fusion architectures that integrate multi-modal battlefield data, AI is reshaping how armed forces plan, target, and execute military operations, ushering in an era where intelligent systems operate alongside, and under the supervision of, human decision-makers (Hoffberger-Pippan, E., & Dahlmann, 2024; Raska, 2024).

Nevertheless, as AI assumes an ever-greater role in various critical decision-making activities, it becomes necessary that these models are designed with both legal and ethical constraints at their core (Panwar, 2024). Compliance with international humanitarian law, human rights principles, and societal expectations demands that AI decision-support systems be transparent, auditable, and aligned with enduring norms such as distinction, proportionality, and necessity. Ethical frameworks further require respect for civilian safety, minimization of unintended harm, and pathways for human oversight to guard against bias, overreach, or malfunction (Lekea, Kraska, 2021; Lekeas & Topalnakos, 2023; Maathuis & Chockalingam, 2023). Embedding these considerations since the design of the AI systems through mechanisms such as value-aligned reward shaping, human-in-the-loop control, and formal verification, ensures that technological development does not outpace moral and legal accountability, thereby preserving the legitimacy of military action under the rule of law.

Among the most sensitive applications of AI in the targeting cycle is the proportionality assessment, a core decision-making process that weighs anticipated military advantage against the estimated collateral damage (Corn & Schoettler, 2015; Dorsey & Bo, 2025). Being defined in the Additional Protocol I to the Geneva Conventions, the principle of proportionality prohibits attacks whose expected incidental damage to civilians or civilian objects would be excessive relative to the concrete and direct military advantage (Dinstein, 2005; Gillard, 2018; Fard & Maathuis, 2021). Implementing this judgment requires dynamic evaluation of uncertain battlefield data such as target value, civilian presence, and weapon effects under time pressure (Li et al.,

2023). In this case, intelligent and adaptive AI solutions could support commanders by rapidly synthesizing sensor feeds, modelling potential collateral effects, and recommending actions that balance operational objectives with humanitarian constraints (Rashid et al., 2023; Roy, 2024). Designing such systems to learn and adapt while remaining firmly anchored in legal-ethical criteria is essential for responsible AI-enabled targeting in contemporary and future conflict settings.

Nevertheless, despite a rich corpus of military-legal, ethical, and technical scholarship on proportionality assessment in military operations, most existing frameworks remain largely prescriptive or descriptive, lacking mechanisms for real-time learning and adaptive decision-making. Specifically, legal analyses and doctrinal studies meticulously articulate the principles of distinction and proportionality, while ethical and philosophical works explore the moral imperatives behind minimizing civilian harm. At the same time, intelligent and computational models ranging from heuristic rule-based systems to probabilistic risk-assessment algorithms offer quantitative methods for estimating collateral effects (Watkin, 2005; Guiora, 2012; Cohen & Zlotogorski, 2021; Maathuis & Chockalingam, 2023; Mirzoev, 2025). Hence, there is a gap in dynamic, self-improving systems capable of refining proportionality judgments through experience, which is critical for ensuring sustained compliance under the uncertainties of modern battlefields. To address this shortfall, we propose a dual-architecture modelling system that implements SARSA and PPO algorithms to operationalize proportionality assessment as an adaptive and hybrid process (Bonnici et al., 2021; Kamble et al., 2025). The system defines each engagement scenario as a discrete state vector of collateral damage and military advantage parameters and trains two agents under divergent rule semantics. Specifically, In Case 1, the SARSA and PPO agents learn proportionality under a strict collateral damage paradigm that excludes psychological injury, thereby encoding traditional kinetic effects into their decision policies. At the same time, in Case 2, the state definition is enriched to encompass psychological harm as a component of collateral damage, challenging the agents to internalize an expanded ethical calculus. By comparing on-policy temporal-difference learning (SARSA) with clipped policy-gradient optimization (PPO) across these two cases, the system demonstrates how reinforcement-learning architectures can both respect established legal norms while adapting to more nuanced ethical considerations in targeting decision-making support.

The outline of this article is structured as follows. In the second section, an overview of the background required in technical terms is provided together with a discussion on related studies carried out in the military domain. Section 3 presents the methodological approach followed to achieve the aim of this research. Section 4 shows the implementation choices made to build the two models proposed in this research. Section 5 discusses the evaluation process and the evaluation results obtained for both cases considered. Conclusively, in Section 6 are discussed reflective points and future research perspectives.

2. Research Background

SARSA and PPO are two prominent RL algorithms offering distinct mechanisms for sequential decision-making. SARSA is an on-policy, model-free control algorithm in which the learning agent updates its state-action value function (Q-value) based not only on the current experience, but also on the subsequent action chosen under its current policy. In this process, at each step, the agent observes its current state, selects an action (often using an epsilon-greedy policy), receives a reward, moves to a new state, selects a new action, and updates its Q-table with this information. This iterative learning process enables SARSA to optimize its decision policy by continually refining value estimates toward maximizing cumulative rewards through balancing exploration and exploitation. Furthermore, the on-policy nature ensures that SARSA's updates reflect the policy currently used by the agent, making it particularly responsive to the consequences of the agent's actual trajectory through the state space (Rummery & Niranjan, 1994; Qiang & Zhongli, 2021; Yao et al., 2025). In contrast, PPO represents an advanced family of policy gradient methods designed to improve the stability and efficiency of policy-based RL training, especially in high-dimensional or continuous action spaces. It is characterized by the use of a stochastic policy, direct policy optimization, and its core innovation which represents a clipped surrogate objective that constrains the magnitude of policy updates to prevent destabilizing shifts. In this process, the agent alternates between sampling trajectories from its current policy, estimating advantages, and updating its policy parameters in a direction that incrementally improves expected performance while ensuring proximal steps to the existing policy. In practical terms, PPO leverages both an actor (policy function) and a critic (value function), employing parallel updates and batch learning to stabilize convergence. This strategic combination allows PPO to maintain robust learning even in environments where reward signals are sparse, delayed, or noisy which are circumstances that can confound traditional value-based algorithms (Schuman et al., 2017; Zeng et al., 2024).

The key elements distinguishing SARSA and PPO relate to their learning mechanisms and adaptability. Specifically, SARSA is relying on explicit state-action evaluations via Q-tables suits smaller, discrete environments but may limit scalability. Its on-policy feature provides stability in stochastic settings. Nevertheless, it is less effective when exploration is insufficient in complex state spaces. At the same time, PPO is valued for its versatility, scalability, and conservative updates, which collectively produce reliable policy improvement in both discrete and continuous domains. Together, SARSA and PPO exemplify the evolution of RL algorithms toward broader applicability, stability, and real-world applications and impact across robotics, autonomous systems, and adaptive decision-support domains (Naeem, Rizvi & Coronato, 2020).

In the military domain, deep RL has been successfully applied to autonomous UAV navigation, where quadcopters learn to navigate through obstacle-laden terrain using only onboard sensors and trial-and-error exploration (Fagundes-Junior et al., 2024). At the same time, hierarchical RL techniques were employed for RF-signal-based UAV detection, training agents to discriminate friend from foe by processing radio-frequency signatures in real time (Ahirrao, Yadav & Kumar, 2024). In search-and-rescue (SAR) operations, RL agents using received-signal-strength indicators learn to localize trapped personnel in GPS-denied indoor environments, by this reducing victim-location time. Furthermore, multi-agent deep RL is used in various UAV swarm surveillance, coordinating dozens of drones to cover and track ground targets collaboratively and RL-driven path-planning algorithms optimize trajectories for reconnaissance missions in contested airspaces (Lyu et al., 2023).

In guided SARSA UAV (Unmanned Aerial Vehicle) path-planning, agents learn to avoid dynamic threats while minimizing fuel consumption, outperforming conventional Q-learning in convergence speed. Here the model underpins network-lifetime maximization in multi-UAV communication networks, where agents schedule flight paths to reduce information age metrics and ensure seamless battlefield connectivity (Luo et al., 2018). In the same context, X propose a SARSA-based protocol to optimize data-collection points for rotary-wing UAVs, balancing channel conditions and mission deadlines to maximize throughput (Joshi, Kalita & Gurusamy, 2023). In a multi-agent setting, for allocation purposes, each UAV agent independently learns bandwidth allocation strategies to support distributed sensing and communication. At the same time, an enhanced PPO algorithm was developed for intelligent joint operations planning, integrating priority sampling and constrained loss functions to optimize C2 decisions in simulation platforms (He et al., 2019). In high-speed intercept guidance, PPO policies learn robust manoeuvre laws to engage agile, manoeuvring targets under kinematic constraints. In the context of threat evaluation, Eval-PPO is deployed for transforming multi-dimensional enemy attributes into a unified RL reward signal for precise, context-aware threat assessments (Sun et al., 2025).

Hence, the deployment of RL algorithms such as SARSA and PPO in military applications offers significant advantages for autonomy, adaptability, and decision-making support. SARSA's on-policy updates facilitate rapid learning in structured, low-noise environments, enabling real-time adaptation to evolving mission constraints with minimal offline computation. At the same time, PPO's clipped gradient updates and decoupled value estimates yield robust convergence even in high-dimensional, stochastic settings, ensuring that learned policies adhere to safety and legal constraints under partial observability. These methods can integrate both domain knowledge and field data as reward shaping or policy priors, providing transparent, auditable decision traces. Taking into account the fact that military operations continue to demand high-throughput, low-latency decision loops, RL represents a versatile toolkit for developing responsible AI that balances mission effectiveness with adherence to humanitarian and legal and ethical norms and principles.

3. Research Methodology

This research adopts the Design Science Research (DSR) methodology as framed (Kuechler & Vaishnavi, 2012; Peffers, Tuunanen & Niehaves, 2018) in order to build two models based on SARSA and PPO reinforcement learning algorithms to conduct the proportionality assessment in military operations. The DSR process initiated with a problem identification phase, recognizing the need to build responsible and trustworthy AI-based tools that can encode and model complex proportionality rules and constraints capturing both the standard and a more ethically-compliant perspective on collateral damage which also includes psychological harm (Maathuis, Pieters & Van Den Berg, 2018). Subsequently, the objectives of the artefact were formulated to implement and systematically optimize the SARSA and PPO models, each representing distinct reinforcement learning paradigm: SARSA as an on-policy temporal-difference learner and PPO as a policy-gradient actor-critic method. The design phase implied constructing these decision-support models, carefully selecting hyperparameters to balance learning speed, stability, and generalization capacity across deterministic rule sets of varying

complexity. From there, an Iterative evaluation cycle involved quantitative performance assessment through reward trajectories, classification accuracy, and confusion matrices reflecting the faithful reproduction of proportionality decisions.

In alignment with DSR's emphasis on rigor and relevance, the evaluation incorporated extensive simulation experiments comparing both algorithms under controlled scenarios representative of operational decision spaces. In this process, metrics such as convergence speed, reward stability, and sensitivity to hyperparameter variation were systematically analysed to understand each artifact's strengths and limitations. Insights gained from these analyses contributed to design knowledge, informing recommended strategies for hyperparameter calibration, such as conservative learning rates and exploration parameters to mitigate SARSA's brittleness or PPO's conservative policy updates that enhance robustness. Furthermore, the results of the comparative analysis are presented to address the audience and further recommendations are provided for further development and deployment. This methodical and iterative development process, grounded in DSR principles, ensures that the resultant SARSA and PPO models contribute to ongoing effort dedicated to building responsible and trustworthy military AI systems that account multidimensional operational and socio-technical domain aspects.

4. Implementation of Models

The proportionality assessment is formulated as a discrete Markov decision process (MDP) whose state space encodes the critical factors governing lawful use of force. Each state s=(s1,s2,s3,s4) is a four-tuple: collateral injury severity (s1), collateral death severity (s2), object damage presence (s3), and military advantage magnitude (s4s). In Case 1, s1 quantifies only physical injury (Low/Medium/High), whereas in Case 2, s1 additionally incorporates psychological harm in the collateral damage component (Maathuis, Pieters & Van den Berg, 2018b). The action set A={Proportional, Disproportional} reflects the binary judgment mandated by Article 51(5)(b) of Additional Protocol I. In this process, a tabular value function or policy is learned over these 54 discrete states, ensuring transparent traceability from scenario features to targeting decisions.

The first model first implements the on-policy algorithm to estimate Q(s,a), the expected return of executing action aa in state ss under the current policy. In this process, the initialization Q(s,a)=0 takes place and a further iteration over uniformly sampled states is carried out. At each step, the agent selects action aa according to an ϵ-greedy policy with values ϵ=0.1 for Case 1 and ϵ=0.2 for Case 2 in order to ensure sufficient exploration. Upon observing reward r∈{+10,−10} for correct or incorrect classification, and next state s', the update

$$Q(s,a) \;\leftarrow\; Q(s,a) + \alpha\big[r + \gamma\, Q(s',a') - Q(s,a)\big]$$

is applied, where a' is the subsequent action chosen by the same ϵ-greedy policy. Further, α=0.1\alpha=0.1 in Case 1 is set for rapid learning and α=0.01\alpha=0.01 is set in Case 2 in order to accommodate the added complexity of psychological injury. Here, a discount factor γ=0.9 balances immediate and future rewards, encapsulating both instant proportionality judgments and longer-term consistency across engagements.

The second model adapts PPO to the tabular MDP, maintaining separate policy parameters π(s,a) and state-value estimates V(s). Here, each state s is flattened to a unique index i, enabling storage of a softmax policy vector πi and scalar V(i). During training, the actions are sampled from πi and advantages A=r−V(i) are computed. Then the policy parameters are updated via the clipped surrogate objective

$$L_{\text{clip}} = \mathbb{E}\Big[\min\big(\rho(\theta)\,A,\ \text{clip}(\rho(\theta), 1-\epsilon_c, 1+\epsilon_c)\,A\big)\Big]$$

where ρ is the likelihood ratio and ϵc=0.1 limits the magnitude update. At the same time, V(i) is regressed toward observed returns with learning rate 0.01. Here, a policy learning rate of 0.001 and a discount factor of 0.9 ensure smooth, monotonic improvement without destabilizing large updates, allowing the policy to generalize seamlessly from Case 1 to Case 2.

Hence, although both SARSA and PPO operate on the same state–action schema, their update mechanics drive different learning dynamics across the two cases. SARSA's bootstrapped, on-policy nature leads to rapid reward gains when only physical damage is considered as part of the collateral damage component, but requires careful tuning (α↓, ϵ↑) to avoid premature convergence in the richer Case 2. PPO's clipped policy updates, by contrast, naturally accommodate increased rule complexity without hyperparameter changes,

providing stable convergence and principled control over update size. This implies that both embed the legal and ethical core aspects of the proportionality assessment deriving rewards from compliance with IHL's harm vs advantage calculus, thus ensuring that learned policies are not merely accurate, but also aligned with humanitarian norms.

5. Evaluation

The SARSA agent was evaluated on both Case 1 (physical collateral damage only) and Case 2 (physical and psychological damage) using three complementary metrics: reward-per-episode convergence, confusion-matrix classification accuracy, and precision/recall breakdown. In Case 1 ($\alpha = 0.1$, $\varepsilon = 0.1$), the reward curve plateaued by approximately 100 episodes at a mean total reward of +48, indicating rapid policy stabilization. However, the confusion matrix revealed only about 44 % accuracy which means that SARSA correctly classified a majority of Proportional states, but misclassified over half of the Disproportional scenarios. This shows a bias toward high-frequency positive outcomes under limited exploration. In Case 2 ($\alpha = 0.01$, $\varepsilon = 0.2$), learning slowed (rewards stabilized around 200 episodes), yet accuracy declined further to about 41 %, with persistent false-positives in cases involving high psychological injury (Figure 1). These results shows SARSA's susceptibility to bootstrapping bias and its difficulty in covering a richer state space despite extended exploration.

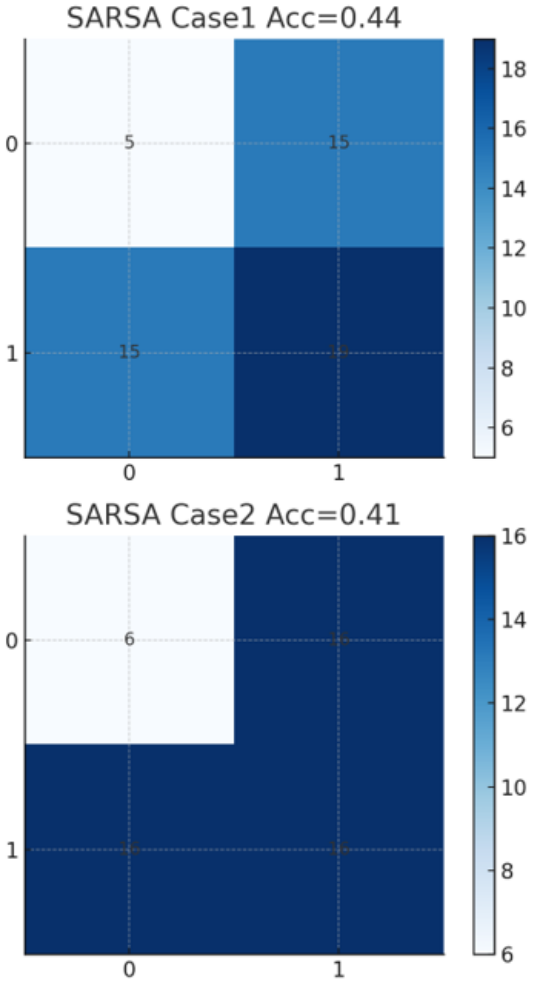


Figure 1: SARSA's accuracy and confusion matrix

The PPO agent demonstrated different dynamics. For Case 1 (policy_lr = 0.001, value_lr = 0.01, clip = 0.1), the reward curve rose more gradually stabilizing near +52 around 600 episodes. Nevertheless, it had a perfect (100 %) classification accuracy, as confirmed by a confusion matrix with zero false-positives or negatives. Precision and recall for both P and DP reached unity, underscoring the model's balanced performance. In Case 2, PPO maintained the same hyperparameters and achieved convergence of reward (~+50) by ~800 episodes, again delivering 100 % accuracy (Figure 2). The stability of reward trajectories, evidenced by low

variance beyond the initial exploration phase, demonstrates PPO's robust policy improvements via its clipped surrogate objective and decoupled value function, even when state semantics are expanded to include psychological harm.

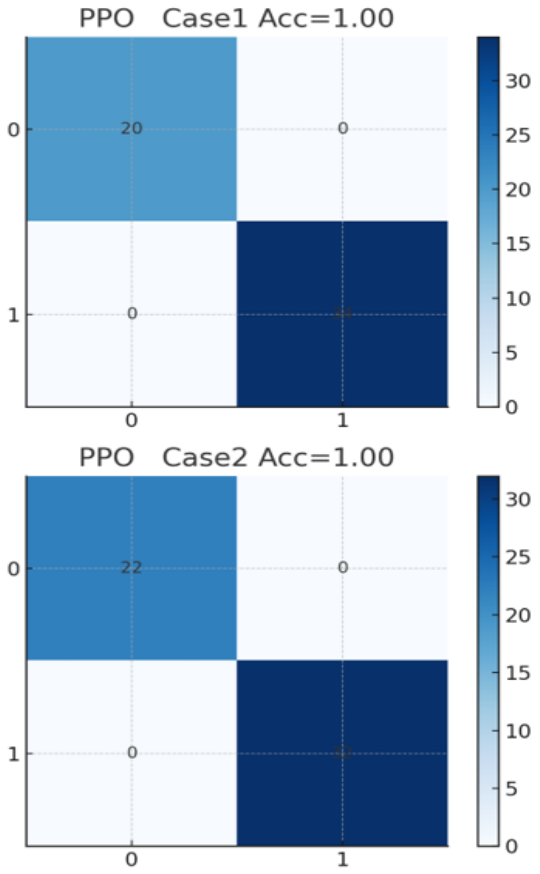


Figure 2: PPO's accuracy and confusion matrix

Further, a comparative analysis is conducted for both models and on both cases. Accordingly, for Case 1, SARSA outpaces PPO in learning speed by reaching its reward plateau in ~100 episodes vs. PPO's ~600, but at the cost of classification fidelity (44 % vs. 100 %). Furthermore, SARSA's rapid bootstrap updates allow quick acquisition of high-frequency state–action values, but lead to overfitting on Proportional outcomes and under-exploration of minority DP states. At the same time, PPO's conservative policy gradients require more iterations to learn, but systematically refine action probabilities across all 54 scenarios, ensuring both high precision and recall. Hence, while SARSA may be advantageous in time constrained settings with simple mappings, PPO offers superior reliability for critical proportionality judgments. For Case 2, SARSA's limitations become more pronounced: even with increased exploration ($\varepsilon = 0.2$) and reduced learning rate ($\alpha = 0.01$), its reward plateau does not translate to improved accuracy, and it continues to misclassify nuanced Disproportional scenarios. At the same time, PPO, maintaining fixed hyperparameters, seamlessly generalizes to the expanded state space, with reward convergence around 800 episodes and 100 % accuracy. This contrast shows PPO's robustness to rule complexity and its ability to integrate evolving ethical dimensions without hyperparameter re-engineering (Figure 3). For operational deployments requiring both adaptability and uncompromising legal compliance, PPO is then seen as the preferable modelling architecture. Nevertheless, given the existence of a potential overfitting as a generalization issue, additional considerations could be given in the future to the inclusion of other relevant variables and dataset for both design and evaluation purposes, and to be able to conduct a comprehensive analytical evaluation.

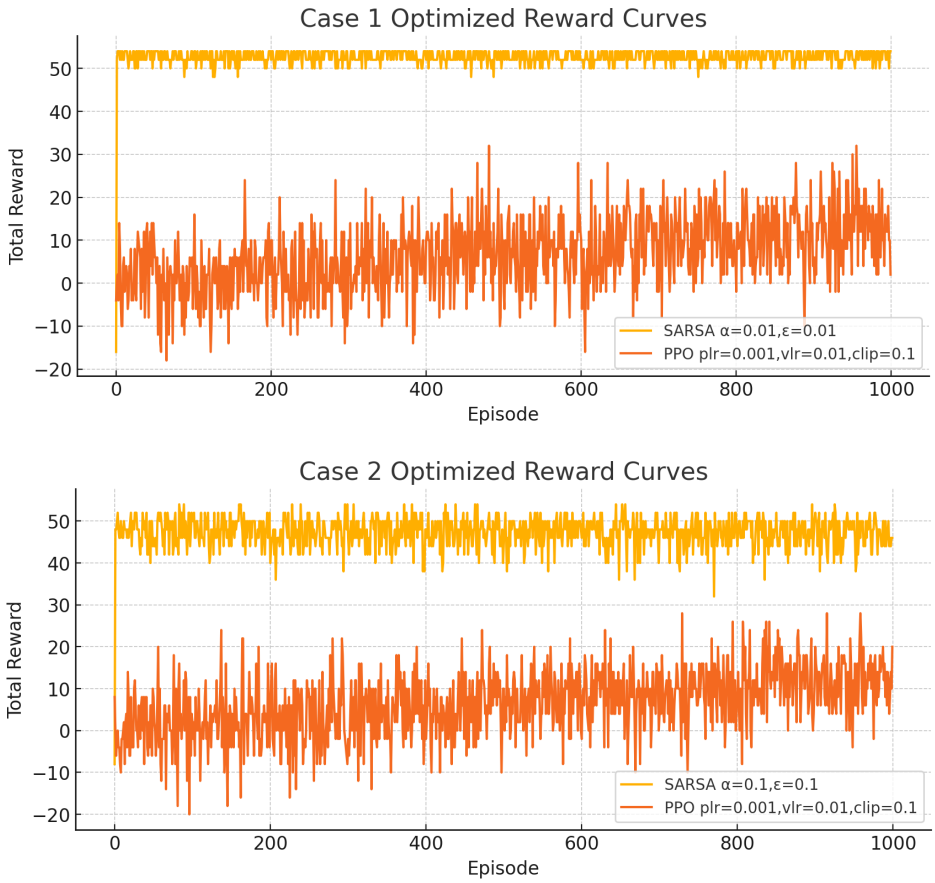


Figure 3: Rewards for both SARSA and PPO models

6. Conclusions

This research proposes two reinforcement learning modelling approaches using the SARSA and PPO algorithms to conduct the proportionality assessment in military operations. It does that following a DSR methodological approach and considering a comparative analysis for evaluation purposes between the two modelling approaches considered applied to the two modalities of framing collateral damage inside the proportionality assessment process considered. Based on the comparative analysis conducted in this research, a series of differences in their learning dynamics, robustness, and suitability for complex legal and ethical rule and constraints encoding is observed. Specifically, SARSA exhibits rapid initial convergence in simpler state spaces, as evidenced by its swift reward plateau in Case 1 further showing challenges in deterministic but high-dimensional decision environments where incomplete exploration and reliance on immediate next-action updates lead to persistent errors. In contrast, the PPO-style approach demonstrates superior stability and accuracy across both use cases, ultimately achieving perfect classification with slower reward accumulation and requiring several hundred episodes to stabilize. Furthermore, its robustness makes PPO more adept at encoding explicit proportionality constraints within deterministic frameworks, providing reliable decision support in contexts demanding high ethical and operational certainty. From a practical standpoint, PPO's trade-off such as increased training time for enhanced global policy refinement and reduced error propagation is positioning it as a more suitable candidate for complex military decision-making scenarios, whereas SARSA's sample efficiency may better serve simpler or more stochastic environments where rapid and less stable learning is permissible.

Future research directions will focus on integrating function approximation techniques such as deep neural networks within both SARSA and PPO frameworks in order to extend their applicability to continuous and high-dimensional collateral damage metrics commonly encountered in real-world military contexts and settings. This advancement could enable scalable generalization beyond tabular representations, addressing current limitations related to state space explosion. Another future research perspective involves incorporating memory-enhanced architectures, like recurrent neural networks, or belief-state estimators to tackle partial observability inherent in operational environments where damage assessments and target states are often

incomplete or uncertain. Additionally, novel training regimes employing reward shaping and curriculum learning where the rule complexity is progressively increased, could mitigate instability issues and accelerate convergence, particularly for SARSA-based methods. Together, these directions aim to enhance the efficiency, robustness, and ethical reliability of AI-based proportionality assessment systems deployed in dynamically evolving military operations.

Declaration: For this research, no ethical clearance is required and no AI tools were used in the creation of this article.

References

Ahirrao, Y. V., Yadav, R. P., & Kumar, S. (2024). RF based UAV detection and identification enhanced by machine learning approach. *IEEE Access*.

Bonnici, R. S., Saliba, C., Caligari, G. E., & Bugeja, M. (2021). Exploring reinforcement learning: A case study applied to the popular snake game. In *The International Conference on Intelligent Systems & Networks* (pp. 169-192). Cham: Springer International Publishing.

Chen, Y., Ji, C., Cai, Y., Yan, T., & Su, B. (2024). Deep reinforcement learning in autonomous car path planning and control: A survey. *arXiv preprint arXiv:2404.00340*.

Cohen, A., & Zlotogorski, D. (2021). *Proportionality in international humanitarian law: Consequences, precautions, and procedures*. Oxford University Press.

Cole, A., Howard, D., Latiff, R., Lucas, G., Reichberg, G. M., & Roy, K. (2024). Artificial intelligence in strategic planning and military operations. In *Cyber security in the Age of artificial intelligence and autonomous weapons* (pp. 120-133). CRC Press.

Corn, G., & Schoettler Jr, J. A. (2015). Targeting and civilian risk mitigation: The essential role of precautionary measures. *Mil. L. Rev.*, *223*, 785.

Dinstein, Y. (2005). Collateral damage and the principle of proportionality. In *New Wars, New Laws? Applying Laws of War in 21st Century Conflicts* (pp. 211-224). Brill Nijhoff.

Dorsey, J., & Bo, M. (2025). AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human (e) Dimension.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Fagundes-Junior, L. A., de Carvalho, K. B., Ferreira, R. S., & Brandão, A. S. (2024). Machine learning for unmanned aerial vehicles navigation: An overview. *SN Computer Science*, *5*(2), 256.

Gillard, E. C. (2018). Some Reflections on the Incidental Harm Side of Proportionality Assessments. *Vand. J. Transnat'l L.*, *51*, 827.

Govinda, S., Brik, B., & Harous, S. (2025). A Survey on Deep Reinforcement Learning Applications in Autonomous Systems: Applications, Open Challenges, and Future Directions. *IEEE Transactions on Intelligent Transportation Systems*.

Guiora, A. N. (2012). Targeted killing: When proportionality gets all out of proportion. *Case W. Res. J. Int'l L.*, *45*, 235.

He, H., Wang, W., Zhu, Y., Li, X., & Wang, T. (2019). An operation planning generation and optimization method for the new intelligent combat SoS. *IEEE Access*, *7*, 156834-156847.

Hoffberger-Pippan, E., & Dahlmann, A. (2024). Digital battlefield: concept, technology and prospects. In *Research Handbook on Warfare and Artificial Intelligence* (pp. 76-98). Edward Elgar Publishing.

Joshi, P., Kalita, A., & Gurusamy, M. (2023). Reliable and efficient data collection in UAV-based IoT networks. *arXiv preprint arXiv:2311.05303*.

Kamble, A. R., Jiet, M. M., Weerarathna, I. N., Dhande, S. A., Izankar, S. V., & Puri, C. G. (2025, March). Reinforcement learning model, algorithms, and its application. In *AIP Conference Proceedings* (Vol. 3227, No. 1, p. 050012). AIP Publishing LLC.

Kraska, J. (2021). Command Accountability for AI Weapon Systems in the Law of Armed Conflict. *International Law Studies*, *97*(1), 22.

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information systems*, *13*(6), 3.

Lekea, I., Lekeas, G., & Topalnakos, P. (2023). Exploring Enhanced Military Ethics and Legal Compliance through Automated Insights: An Experiment on Military Decision-making in Extremis. *Conatus-Journal of Philosophy*, *8*(2), 345-372.

Li, G., He, G., Zheng, M., & Zheng, A. (2023). Uncertain sensor–weapon–target allocation problem based on uncertainty theory. *Symmetry*, *15*(1), 176.

Luo, W., Tang, Q., Fu, C., & Eberhard, P. (2018, June). Deep-sarsa based multi-UAV path planning and obstacle avoidance in a dynamic environment. In *International Conference on Swarm Intelligence* (pp. 102-111). Cham: Springer International Publishing.

Lyu, M., Zhao, Y., Huang, C., & Huang, H. (2023). Unmanned aerial vehicles for search and rescue: A survey. *Remote Sensing*, *15*(13), 3266.

Maathuis, C. (2022). An Outlook of Digital Twins in Offensive Military Cyber Operations. In *European Conference on the Impact of Artificial Intelligence and Robotics* (Vol. 4, No. 1, pp. 45-53).

Maathuis, C., Pieters, W., & Van Den Berg, J. (2018). A computational ontology for cyber operations. In *Proceedings of the 17th European Conference on Cyber Warfare and Security* (pp. 278-288).

Maathuis, C., Pieters, W., & van den Berg, J. (2018b). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C., & Chockalingam, S. (2023). Modelling the influential factors embedded in the proportionality assessment in military operations. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 218-226).

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., ... & Morimoto, J. (2022). Deep learning, reinforcement learning, and world models. *Neural Networks*, *152*, 267-275.

Mirzoev, R. (2025). The Principle of Proportionality in International Humanitarian Law: Evolution, Application, and Accountability in Modern Armed Conflicts.

Naeem, M., Rizvi, S. T. H., & Coronato, A. (2020). A gentle introduction to reinforcement learning and its application in different fields. *IEEE access*, *8*, 209320-209344.

Panwar, R. S. (2024). Artificial intelligence in military operations: technology, ethics and the Indian perspective. In *Artificial Intelligence, Ethics and the Future of Warfare* (pp. 216-226). Routledge India.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Probasco, E., Toner, H., Burtell, M., & Rudner, T. G. (2025). AI for Military decision-making. *Center for Security and Emerging Technology*, 5-26

Qiang, W., & Zhongli, Z. (2011). Reinforcement learning model, algorithms and its application. In *2011 International Conference on Mechatronic Science, Electric Engineering and Computer (MEC)* (pp. 1143-1146). IEEE.

Rashid, A. B., Kausik, A. K., Al Hassan Sunny, A., & Bappy, M. H. (2023). Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International journal of intelligent systems*, *2023*(1), 8676366.

Raska, M. (2024). Reimagining Defense Innovation: Defense AI in Singapore. In *The Very Long Game: 25 Case Studies on the Global State of Defense AI* (pp. 555-580). Cham: Springer Nature Switzerland.

Roy, K. (Ed.). (2024). *Artificial intelligence, ethics and the future of warfare: Global perspectives*. Taylor & Francis.

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems* (Vol. 37, p. 14). Cambridge, UK: University of Cambridge, Department of Engineering.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Sun, W., Li, S., Zou, Q., & Liao, Z. (2025). Eval-PPO: Building an Efficient Threat Evaluator Using Proximal Policy Optimization. *arXiv preprint arXiv:2503.12098*.

Taye, M. M. (2023). Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, *12*(5), 91.

Watkin, K. (2005). Assessing proportionality: moral complexity and legal rules. *Yearbook of International Humanitarian Law*, *8*, 3-53.

Yao, G., Zhang, N., Duan, Z., & Tian, C. (2025). Improved SARSA and DQN algorithms for reinforcement learning. *Theoretical computer science*, *1027*, 115025.

Zeng, Y., Cai, R., Sun, F., Huang, L., & Hao, Z. (2024). A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems*.

Genetic Algorithms for Proportionality Assessment in Military Operations

Clara Maathuis
Open University of the Netherlands, Heerlen, The Netherlands

clara.maathuis@ou.nl

Abstract: This research proposes the application of Genetic Algorithms (GAs) as an effective computational approach for proportionality assessment in military operations. With this scope, two distinct GA-based models are proposed: one evaluating proportionality excluding psychological damage from collateral damage considerations, and another integrating psychological damage into the collateral damage component of this assessment. Each model encodes the operational assessment rules as binary classification policies, mapping multidimensional states defined by levels of injury, death, object damage, and military advantage to proportional or disproportional decisions. Through evolutionary optimization involving selection, crossover, mutation, and fitness evaluation across generation, the GAs search for classification rules that maximize alignment with expert-defined proportionality judgments. From the evaluation conducted which includes accuracy, fitness progression, population diversity, and confusion matrices, it is seen that the models converge reliably to high-performance solutions, achieving high classification accuracy within the deterministic rule sets. Further, the influence of psychological damage is assessed in relation to the behaviour convergence and classification outcomes. The results show the utility of GAs in automating this military decision-making process as a responsible, transparent, adaptive, and interpretable mechanism for proportionality assessment that is able to incorporate both legal and ethical considerations in operational environments.

Keywords: Military operations, Proportionality, Artificial intelligence, Genetic algorithms, Optimization

1. Introduction

Motto: *"The world is in greater peril from those who tolerate or encourage evil than from those who actually commit it ." (Albert Einstein)*

Recent advances in the Artificial Intelligence (AI) domain are revolutionizing military capabilities across all operational domains (Rashid et al., 2023). Autonomous ground vehicles now traverse unstructured terrain using simultaneous localization and mapping (SLAM) algorithms, while sensor-to-shooter loops leverage convolutional and transformer-based networks for rapid target detection and tracking (Min et al., 2024; Kabir et al., 2024). At the same time, swarm robotics applications integrate decentralized multi-agent coordination protocols, enabling distributed teams of unmanned aerial and maritime systems to execute complex reconnaissance, electronic warfare, and counter-insurgency tasks under contested communications (Abro et al., 2024; Alqudsi & Makaraci, 2025). Meanwhile, the integration of digital twins and synthetic training environments has accelerated AI model validation through realistic scenario replay, reducing field-trial risks and speeding deployment cycles (Maathuis, 2022; Mahajan et al., 2024; Sajadieh & Noh, 2025). These technological strides, coupled with advances in the area of AI security and Explainable AI (XAI), position AI as a transformative force in the modern battlespace.

Nevertheless, the promise of AI-enabled decision-support comes with profound legal and ethical imperatives that must guide system design from the outset. International humanitarian Law (IHL) mandates distinction, proportionality, and precautions to be taken, requiring that autonomous or semi-autonomous systems discriminate combatants from civilians, weigh military advantage against collateral damage, and implement feasible mitigation measures (Egeland, 2016; Cohen, A., & Zlotogorski, 2021; Saxon, 2021). Ethical frameworks such as the IEEE's Ethically Aligned Design and NATO's Emerging and Disruptive Technologies principles contain requirements for algorithmic transparency, bias mitigation, and human-in-the-loop oversight to preserve accountability and public trust (Ray, 2023; Bur & Leslie, 2023; Maathuis & Chockalingam, 2022). Furthermore, AI governance models argue for continuous validation, auditability of decision logs, and robust fail-safe mechanisms in order to prevent unintended escalation or misuse of the AI systems developed and deployed (Sayles, 2024). Embedding these legal and ethical constraints into reward functions, policy priors, and system architectures ensures compliance with societal norms and upholds the moral legitimacy of military AI applications.

At the core of targeting decision-support lies the principle of proportionality, one of the most consequential judgments in the engagement cycle. Under Article 51(5)(b) of Additional Protocol I, commanders must refrain from attacks whose expected incidental collateral damage on civilians and civilian objects is excessive relative to the concrete and direct military advantage (Bellal & Casey-Maslen, 2022; Maathuis & Chockalingam, 2023).

Operationalizing this requirement demands AI solutions that synthesize heterogeneous data through sources such as electro-optical imagery, human intelligence reports, and probabilistic collateral-damage models under uncertainty and time pressure (Jafariandehkordi,, 2024). To this end, adaptive Reinforcement Learning (RL) and evolutionary computational techniques can learn from after-action reviews, refining estimations of non-kinetic effects such as infrastructure disruption and psychological impacts (Gautam, 2023; Lin et al., 2025; Fard & Maathuis, 2021). By continuously updating proportionality metrics and providing transparent rationales for engagement recommendations, these intelligent systems can enhance human decision-making while rigorously adhering to both legal mandates and humanitarian values.

While various research and practitioner efforts were dedicated to understand, capture, and model various dimensions of the proportionality assessment and the assessment itself, most contributions do not account a continuous refinement mechanism. While legal, ethical, and philosophical efforts delineated the contours of permissible force under IHL (Boutin, B., & Woodcock, 2024; Capone, 2025) and analyses further expanded the discourse to include concepts such as collateral psychological harm and infrastructure resilience, other efforts have translated these norms into threshold heuristics or weighted scoring models. Nevertheless, a limited number of studies incorporate adaptive learning or feedback loops that enable a system to evolve its proportionality judgments in response to operational outcomes or shifting tactical contexts. As a result, existing methods often deal with various challenges when dealing with novel scenarios or ambiguous evidence, underscoring the need for models that can self-improve and generalize beyond their initial rule sets. To address this gap, a Genetic Algorithms-based model (Pétrowski, A., & Ben-Hamida, 2017; Gerges et al., 2018) is proposed in this research for conducting the proportionality assessment in military operations as an evolutionary search for optimal classification policies. In this way, each candidate policy is encoded as a binary chromosome over the 54 possible engagement states that evolves through selection, crossover, and mutation to maximize fidelity to the proportionality rules. From here, the system is evaluated under two distinct paradigms. In the first case, collateral damage is narrowly defined by physical injury, death, and object destruction, reflecting a traditional kinetic perspective. And in the second case, the damage vector is expanded in order to include psychological harm as well, thereby embedding a more expansive humanitarian calculus. By comparing convergence dynamics, classification accuracy, and policy robustness across these two cases, the GAs system proposed demonstrates both the feasibility of adaptive proportionality modelling and the capacity to integrate evolving ethical considerations into responsible AI-enabled targeting decision-making support.

The outline of this article is structured as follows. Section 2 discusses the theoretical background of GAs and a series of relevant related studies conducted in this domain. Section 3 presents the methodological approach taken in this research. Section 4 presents the modelling context and decisions made for both use cases considered in the implementation and evaluation processes. Section 5 discusses the evaluation process and the results obtained for both cases considered. At the end, in Section 6, concluding remarks and future research perspectives are provided.

2. Research Background

Genetic Algorithms are a prominent class of optimization and search heuristics inspired by the biological principles of natural selection and genetic evolution, situated within the broader artificial intelligence paradigm. They operate by simulating the evolutionary process to progressively improve a population of candidate solutions to a given problem over successive generations. This process embodies the principle of survival of the fittest where solutions, encoded as chromosomes consisting of genes, are evaluated according to a fitness function that measures their quality or performance (Back, 1996; Sloss, A. N., & Gustafson, 2020). Through iterative genetic operations such as selection, crossover (recombination), and mutation, GAs explore the solution space by favouring higher-fitness individuals for reproduction and introducing diversity to prevent premature convergence (Pandey, H. M., Chaudhary, A., & Mehrotra, 2014; Felix-Saul et al., 2024). The population-based, stochastic nature of GAs allows them to effectively search complex, multimodal, and nonlinear problem landscapes which are often challenging for traditional deterministic or gradient-based optimization methods.

The core elements of a genetic algorithm include an initial population of encoded candidate solutions, a fitness function to evaluate solution quality, and genetic operators that evolve the population mathematically. Selection mechanisms, such as tournament or roulette wheel selection, probabilistically choose the most fit individuals to parent the next generation. In this process, the crossover operator exchanges genetic material between parents to produce offspring potentially combining advantageous traits, while mutation introduces random alterations to genes, ensuring genetic diversity and exploration of the search space (Alhijawi &

Awajan, 2024). These iterative generations continue until stopping criteria, e.g., convergence of fitness, maximum generations, or computational budget are met. Because GAs function without requiring gradient information or explicit modelling of the problem environment, they are well-suited to optimization tasks where objective functions are discontinuous, nondifferentiable, or noisy.

Further, GAs have various applications due to their flexibility, robustness, and global search capability. They are frequently employed in combinatorial optimization, machine learning model tuning, scheduling, and complex system design (Hamdia, Zhuang & Rabczuk, 2021; Zhang et al., 2023). While classic GAs rely on relatively simple encodings and operators, ongoing research focuses on hybridizing GAs with other AI approaches such as neural networks, reinforcement learning, or multi-objective optimization in order to address high-dimensional, dynamic, or uncertain environments. In supply chain management, GAs have been applied to multi-objective network design by minimizing costs and emissions while maximizing reliability, and through demonstrating significant improvements in closed-loop logistics under demand uncertainty and uncovering Pareto-optimal trade-offs in multi-echelon structure (Wu et al., 2022; Gen & Lin, 2023; Maathuis, Pieters & van den Berg, 2018). In healthcare operations, supplier selection and resource allocation problems have been formulated as multi-objective GA tasks, optimizing cost, quality, and delivery time for critical medical supplies (Mohamed et al., 2023). At the same time, in the context of environmental management during the COVID-19 pandemic, GAs optimized municipal waste collection routes and resource allocation in emergency response, achieving up to 15 % performance gains over traditional heuristics (Anđelić et al., 2021; Li et al., 2021). Furthermore, for innovative product design applications such as NASA's ST5 spacecraft antenna, whose complex shape was evolved entirely by GA to meet stringent radiation-pattern requirements highlighting their versatility in engineering design tasks (Alyammahi, 2023).

In the military domain, GAs have been effectively deployed for autonomous UAV path planning, where safety constraints, dynamic obstacles, and multi-objective goals (e.g., fuel efficiency, stealth) create highly nonconvex search spaces (Pan et al., 2021; Liu, 2022). Early studies solved three-dimensional waypoint routing as a Traveling Salesman Problem, using GAs to avoid radars and terrain hazards, thereby enhancing mission success and reducing pilot workload (Ramasamy et al., 2022). More recent work integrates GA with geometric obstacle-avoidance algorithms (e.g., QuickNav), producing advance mission plans that simultaneously respect no-fly zones and real-time sensor updates, resulting in collision-free trajectories with near-optimal path length (Aizat et al., 2023; Debnath et al., 2024). At the same time, parallel implementations on multicore architectures have further improved computation speed, enabling real-time replanning in contested environments.

Moreover, in the context of multi-UAV mission planning, GAs were used for coordinated task allocation and path optimization as multi-objective evolutionary problems (Bai et al., 2022). A hybrid GA with constraint-satisfaction checks and Pareto front analysis efficiently balances makespan, fuel consumption, and coverage metrics across UAV teams, reflecting scalable performance on complex mission datasets (Alqahtani, H., & Kumar, 2024; Özer, 2024). Moreover, in emergency response scenarios such as coordinated search-and-rescue operations, improved GAs (e.g., ERRT-GA) integrate expert heuristics to escape local optima, showing faster convergence and enhanced task throughput in high-pressure, multi-agent settings (Zhan et al., 2024). These studies show GAs' ability to navigate rugged fitness landscapes inherent to joint-operations planning. Furthermore, GAs are used for sensor placement and coverage path planning for both reconnaissance and structural inspection. In one study, a Biased Random Key Genetic Algorithm (BRKGA) solved a Set-Covering Vehicle Routing Problem (SC-VRP), reducing inspection path length by up to 48 % for large 3D structures, thereby optimizing surveillance efficacy in complex terrain (Huang et al., 2023; Oliveira et al., 2024). Such applications demonstrate GAs' strength in producing high-quality, interpretable solutions for discrete, combinatorial military decision-making processes, making them to be considered valuable tools for evolving the next generation of autonomous and semi-autonomous defence systems.

3. Research Methodology

In this research, the Design Science Research methodology is applied in order to develop, evaluate, and communicate a GA-based model for proportionality assessment in military operations (Kuechler & Vaishnavi, 2012; Peffers, Tuunanen & Niehaves, 2018). The research process started with problem identification and motivation, recognizing the need for building computational models that can encode complex proportionality rules and constraints involving collateral damage and psychosocial injury considerations which serves the base for the data and rules used in this research (Maathuis, Pieters & van den Berg, 2018; Maathuis, 2022b). Following this, clear objectives were set to design and implement a GA-based decision-support artifact capable

of mapping state configurations to proportionality decisions through evolutionary optimization. Furthermore, the design and development phase involved encoding these mappings as chromosomes, defining a fitness function aligned with rule-based correctness, and iteratively applying genetic operators such as selection, crossover, and mutation. The model obtained was then instantiated and tested on representative scenarios of varying complexity to ensure it effectively internalizes target classification constraints.

From this stage, the evaluation phase comprised systematic experimentation to measure the artifact's performance through metrics like classification accuracy, fitness progression, and population diversity. The iterative nature of the GA allowed for continuous monitoring of convergence behaviours and robustness to feature complexity, exemplifying the design-as-a-search-process principle. Insights from these evaluations informed refinements of genetic parameters to balance exploration and exploitation. At the same time, the artifact's contribution to design knowledge was articulated with dual audiences in mind: technology developers aiming for implementable AI solutions, and military decision-makers requiring reliable and interpretable proportionality assessment tools. In addition, documentation and communication of the artifact's design rationale, evaluation results, and limitations complete the DSR cycle, are ensuring that the GA model serves as both a practical tool and a research contribution to building responsible and trustworthy military AI (Maathuis, 2024) decision-support systems.

4. Model Implementation

In this research, the proportionality assessment is conceptualized as a combinatorial optimization challenge solvable via GAs. To this end, each candidate solution, i.e., a chromosome, encodes a complete classification policy over the 54 discrete engagement states defined by collateral-damage severity and military advantage. Furthermore, considerations for both cases are provided as follows. In Case 1, the first gene dimension represents only physical injury (Low/Medium/High) as collateral damage, whereas in Case 2 it encompasses both physical and psychological harm as collateral damage. Each gene allele takes one of two values, P (Proportional) or DP (Disproportional), thus forming a 54-bit binary string which corresponds to the number of distinct engagement states of the model. This representation ensures direct interpretability: a chromosome's complete policy can be decoded state by state to yield actionable targeting recommendations.

The process starts with the initialization of a random population of 50–100 chromosomes. Here, the fitness function quantitatively measures how closely each candidate's state–action mapping matches the expert rule set: +1 for each of the 54 correct classifications, and 0 for mismatches, yielding integer fitness in [0,54]. This dense feedback drives selection pressure toward high-accuracy policies, while preserving genetic diversity in early generations. We track both the best fitness and the average population fitness per generation to monitor convergence behaviour and population health. To evolve solutions, the following GA operators are applied:

- Selection via tournament or roulette-wheel, probabilistically favouring higher-fitness chromosomes yet allowing occasional selection of weaker candidates to prevent premature convergence.
- Crossover, implemented as one-point or uniform which recombines parental bit-strings to produce offspring that blend state decisions across both lineages.
- Mutation, executed with a low bit-flip probability (~0.01) which introduces random gene perturbations, enabling exploration of novel policy variants and maintaining genetic diversity. These operations are iterated for 500–1 000 generations (or until fitness stagnation), dynamically recording the population's diversity index, computed as the mean pairwise Hamming distance between chromosomes, to ensure a balance of exploration and exploitation .

From here, the model is executed for both use cases. Although GA hyperparameters (population size, crossover rate ≈ 0.8, mutation rate ≈ 0.01) remain constant, the expanded semantics of Case 2 typically demand more generations to converge. Upon completion, the best chromosome is extracted and the final accuracy is computed, the confusion matrix is generated to assess misclassification patterns, and a plot for fitness and diversity curves over generations is considered. This comprehensive design facilitates the comparison of GA's capacity to discover robust proportionality policies under both strict kinetic and enriched humanitarian perspectives.

5. Model Evaluation

From the evaluation process, the GA model for Case 1 attained rapid convergence to the proportionality rules, with the best-fitness chromosome reaching the maximum score of 54 (100 % classification accuracy) by generation 120. The fitness trajectory (mean and best over generations) exhibited a steep ascent in early

generations, reflecting effective selection and crossover pressure, followed by a gradual plateau as mutation refined boundary cases. At the same time, the population diversity, measured via mean pairwise Hamming distance, declined sharply until generation 80 which shows consensus on high-fitness genes and stabilized thereafter, confirming that exploitation had overtaken exploration without complete genetic homogenization. On this point, the confusion matrix shows zero false positives or negatives across the 54 scenarios, demonstrating the GA's capacity to fully recover the deterministic rule mapping when only physical collateral damage is considered.

Furthermore, by introducing psychological harm in Case 2 increased rule complexity and expanded the semantic burden on the GA. Here, the best chromosome achieved a fitness of 52 out of 54 (≈96.3 % accuracy) by generation 250, with the remaining mismatches concentrated in Medium psychological-damage states with High military advantage. Here, the fitness curve rose more slowly and with greater variance which reflects a more rugged fitness landscape while diversity persisted at moderate levels through generation 150, indicating prolonged exploration of alternate policy variants before convergence. Moreover, the confusion matrix depicts two misclassified states, underscoring that GA can approximate the enriched ethical calculus, but may require enhanced parameter settings (e.g., larger population or higher mutation) to guarantee full rule coverage.

From the comparative analysis conducted in this research, Case 1's simpler state semantics yielded faster convergence (≈120 vs. 250 generations) and sharper diversity reduction, as the fitness landscape contained a single, easily discoverable global optimum. In constrast, Case 2's enriched state space generated a multimodal landscape, delaying the discovery of near-optimal chromosomes and sustaining higher diversity to avoid premature convergence (Figure 1 and Figure 2). This difference illustrates the sensitivity of GA search efficiency to problem complexity and highlights the need for adaptive genetic operator rates or niching strategies when incorporating additional ethical dimensions into the input.

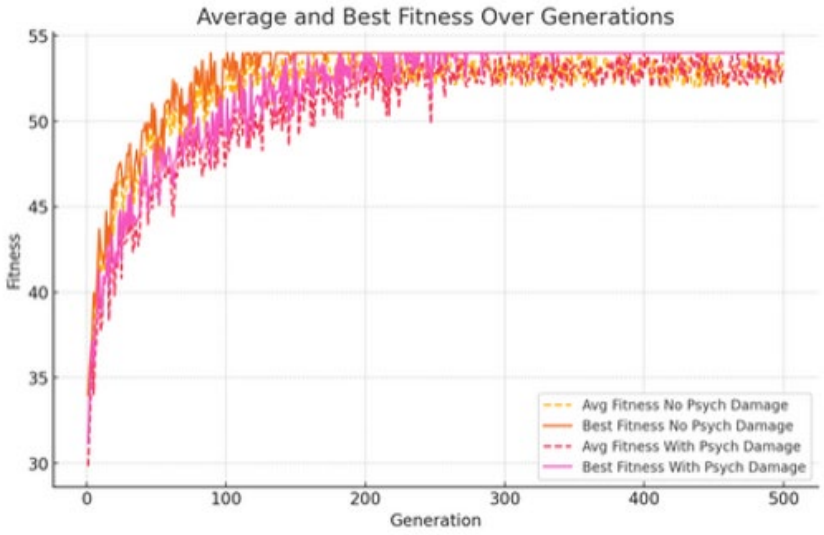


Figure 1: Best fitness over generations

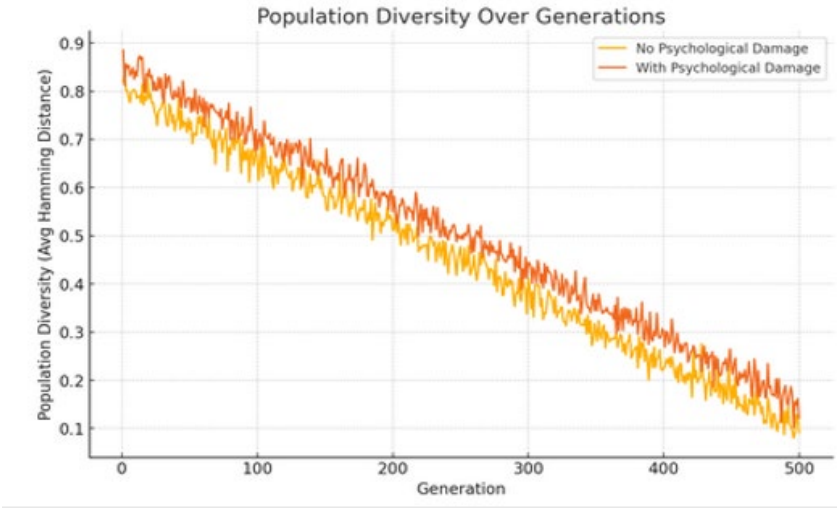


Figure 2: Population diversity over generations

Furthermore, both GA models effectively uncovered high-fidelity policies, Case 2's residual misclassifications point to the limits of fixed GA hyperparameters under more nuanced proportionality rules. The 96.3 % accuracy in Case 2, although high, falls short of the 100 % achieved in Case 1, suggesting that supplementary techniques such as hybridizing GA with local search or increasing mutation rates could bolster robustness (Figure 3 and Figure 4). Nevertheless, the GA framework demonstrates their capability for encoding both strict kinetic and expanded humanitarian perspectives by providing an interpretable, parallel search mechanism that can be tailored to balance convergence speed, diversity retention, and classification accuracy.

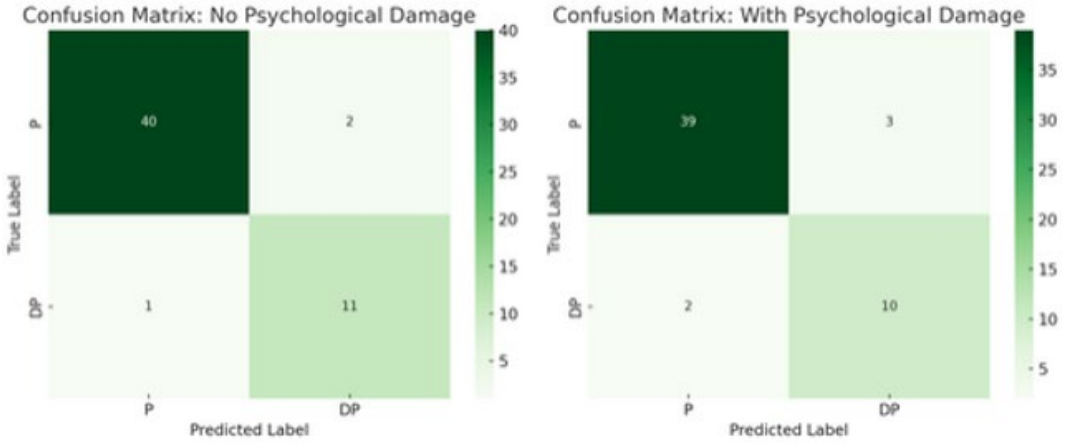


Figure 3: Confusion matrix for both cases

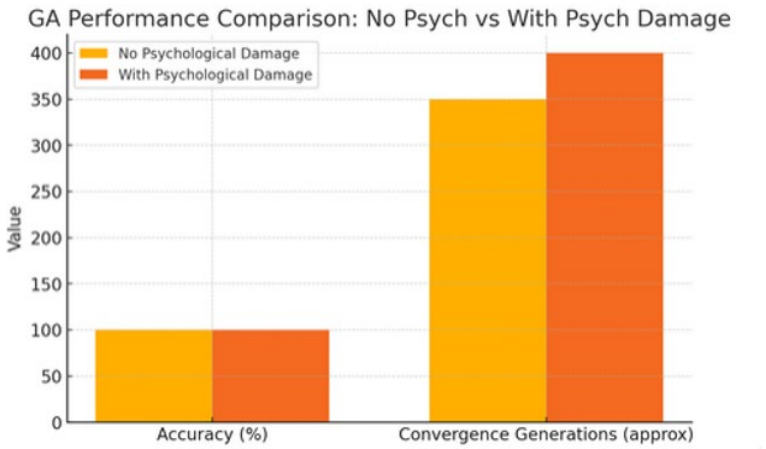


Figure 4: Accuracy and convergence for both cases

6. Conclusions

The application of GA for proportionality assessment in military operations provides a robust evolutionary computation approach capable of discovering effective classification policies for complex legal and ethical decision-making. In this research, this approach encodes candidate solutions as chromosomes representing mappings from states defined by collateral damage and military advantage attributes to proportionality decisions, i.e., being Proportional or Disproportional. The evolutionary process iteratively refines these mappings through fitness-based selection, crossover, and mutation, guided by a fitness function that rewards accurate classification against deterministic rule sets. The experimental implementation conducted in two cases, one excluding psychological damage as part of the collateral damage component and another including it, showed convergence towards near-perfect classification accuracy. In this process, the GA's global search capabilities allow it to explore alternative policy mappings, providing resilience and adaptability in the presence of potential noise or partial rule knowledge, which commonly characterizes real-world military operational contexts and environments. Visualization of fitness curves, population diversity, and confusion matrices demonstrates the GA's progressive learning and convergence stability in both simpler and enriched feature spaces, reinforcing its utility as a decision-support artifact in proportionality assessment.

From an AI and military operational perspective, this application of GAs shows the potential of evolutionary algorithms to balance exploration and exploitation within the challenge of ethically sensitive military targeting decisions. The success in both case scenarios highlights the GA's capacity to handle combinatorially large state-action spaces without requiring gradient information or explicit environment models, a critical advantage in uncertain and dynamic combat settings. Furthermore, the maintenance of population diversity throughout the evolutionary runs safeguards against premature convergence, ensuring robustness as the proportionality constraints increase in complexity with psychosocial injury considerations. From here, future research includes extending the GA framework with multi-objective optimization to integrate additional criteria such as minimizing civilian risk variance and dynamically evolving rules of engagement, which would better reflect operational realities. Another promising direction involves hybridizing GAs with deep learning or reinforcement learning techniques to leverage function approximation for continuous state spaces and enable real-time adaptability in high-stakes battlefield environments. Both the current approach and future extensions aim to enhance the practical applicability and ethical reliability of AI-enabled proportionality assessment systems in future military decision-support platforms to ensure the development and deployment of responsible and trustworthy military AI systems.

Declaration: For this research, no ethical clearance is required and no AI tools were used in the creation of this article.

References

Abro, G. E. M., Ali, Z. A., & Masood, R. J. (2024). Synergistic UAV motion: A comprehensive review on advancing multi-agent coordination. *ICCK Transactions on Sensing, Communication, and Control*, *1*(2), 72-88.

Aizat, M., Qistina, N., & Rahiman, W. (2023). A comprehensive review of recent advances in automated guided vehicle technologies: Dynamic obstacle avoidance in complex environment toward autonomous capability. *IEEE Transactions on Instrumentation and Measurement*, *73*, 1-25.

Alqahtani, H., & Kumar, G. (2024). Efficient routing strategies for electric and flying vehicles: A comprehensive hybrid metaheuristic review. *IEEE Transactions on Intelligent Vehicles*.

Alhijawi, B., & Awajan, A. (2024). Genetic algorithms: Theory, genetic operators, solutions, and applications. *Evolutionary Intelligence*, *17*(3), 1245-1256.

Alqudsi, Y., & Makaraci, M. (2025). UAV swarms: research, challenges, and future directions. *Journal of Engineering and Applied Science*, *72*(1), 12.

Alyammahi, E. R. A. (2023). *Antenna Design Optimization Using Evolutionary and Swarm Intelligence Algorithms* (Doctoral dissertation, Khalifa University of Science).

Anđelić, N., Baressi Šegota, S., Lorencin, I., Mrzljak, V., & Car, Z. (2021). Estimation of COVID-19 epidemic curves using genetic programming algorithm. *Health informatics journal*, *27*(1), 1460458220976728.

Back, T. (1996). *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press.

Bai, H., Fan, T., Niu, Y., & Cui, Z. (2022). Multi-UAV cooperative trajectory planning based on many-objective evolutionary algorithm. *Complex System Modeling and Simulation*, *2*(2), 130-141.

Bellal, A., & Casey-Maslen, S. (2022). *The Additional Protocols to the Geneva Conventions in Context*. Oxford University Press.

Boutin, B., & Woodcock, T. (2024). Aspects of realizing (meaningful) human control: a legal perspective. In *Research Handbook on Warfare and Artificial Intelligence* (pp. 179-196). Edward Elgar Publishing.

Capone, F. (2025). The Centrality of the "Reasonable Military Commander Standard" in the Paradigm Shift. In *Dual-use Objects under International Humanitarian Law: Towards a Paradigm Shift* (pp. 157-194). The Hague: TMC Asser Press.

Cohen, A., & Zlotogorski, D. (2021). *Proportionality in international humanitarian law: Consequences, precautions, and procedures*. Oxford University Press.

Debnath, D., Vanegas, F., Boiteau, S., & Gonzalez, F. (2024). An integrated geometric obstacle avoidance and genetic algorithm tsp model for uav path planning. *Drones*, *8*(7), 302.

Egeland, K. (2016). Lethal autonomous weapon systems under international humanitarian law. *Nordic Journal of International Law*, *85*(2), 89-118.

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Felix-Saul, J. C., García-Valdez, M., Merelo Guervós, J. J., & Castillo, O. (2024). Extending Genetic Algorithms with Biological Life-Cycle Dynamics. *Biomimetics*, *9*(8), 476.

Gautam, M. (2023). Deep Reinforcement learning for resilient power and energy systems: Progress, prospects, and future avenues. *Electricity*, *4*(4), 336-380.

Gen, M., & Lin, L. (2023). Genetic algorithms and their applications. In *Springer handbook of engineering statistics* (pp. 635-674). London: Springer London.

Gerges, F., Zouein, G., & Azar, D. (2018, March). Genetic algorithms with local optima handling to solve sudoku puzzles. In *Proceedings of the 2018 international conference on computing and artificial intelligence* (pp. 19-22).

Hamdia, K. M., Zhuang, X., & Rabczuk, T. (2021). An efficient optimization approach for designing machine learning models based on genetic algorithm. *Neural Computing and Applications*, *33*(6), 1923-1933.

Huang, Y., Chen, B., Lu, W., Jin, Z. X., & Zheng, R. (2022). Asynchronous optimization of part logistics routing problem. *Journal of Global Optimization*, *82*(4), 803-834.

Jafariandehkordi, M. (2024). The AI Battlefield: Legal Challenges of Autonomous Weapon Systems under International Humanitarian Law.

Lin, Y., Lin, F., Cai, G., Chen, H., Zou, L., Liu, Y., & Wu, P. (2025). Evolutionary reinforcement learning: A systematic review and future directions. *Mathematics*, *13*(5), 833.

Liu, J. (2022). An improved genetic algorithm for rapid uav path planning. In *Journal of Physics: Conference Series* (Vol. 2216, No. 1, p. 012035). IOP Publishing.

Kabir, M. M., Jim, J. R., & Istenes, Z. (2024) Fusion-enhanced terrain detection and segmentation for autonomous vehicle navigation: A state-of-the-art systematic review. *Available at SSRN 4797592*.

Kuechler, W., & Vaishnavi, V. (2012). A framework for theory development in design science research: multiple perspectives. *Journal of the Association for Information systems*, *13*(6), 3.

Li, D., Zhou, H., Hui, X., He, X., & Mu, X. (2021). Plasmonic biosensor augmented by a genetic algorithm for ultra-rapid, label-free, and multi-functional detection of COVID-19. *Analytical Chemistry*, *93*(27), 9437-9444.

Maathuis, C., Pieters, W., & van den Berg, J. (2018). A knowledge-based model for assessing the effects of cyber warfare. In *Proceedings of the 12th NATO Conference on Operations Research and Analysis*.

Maathuis, C. (2022). An Outlook of Digital Twins in Offensive Military Cyber Operations. In *European Conference on the Impact of Artificial Intelligence and Robotics* (Vol. 4, No. 1, pp. 45-53).

Maathuis, C. (2022b). On the road to designing responsible AI systems in military cyber operations. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1, pp. 170-177).

Maathuis, C., & Chockalingam, S. (2022). Responsible digital security behaviour: Definition and assessment model. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1).

Maathuis, C., & Chockalingam, S. (2023). Modelling the influential factors embedded in the proportionality assessment in military operations. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 218-226).

Maathuis, C. (2024). Trustworthy Human-Autonomy Teaming for Proportionality Assessment in Military Operations. In *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)* (pp. 1-8). IEEE.

Mahajan, S., Raj, P., & Pandit, A. K. (Eds.). (2024). *Deep Reinforcement Learning and Its Industrial Use Cases: AI for Real-World Applications*. John Wiley & Sons.

Min, C., Si, S., Wang, X., Xue, H., Jiang, W., Liu, Y., ... & Li, X. (2024). Autonomous driving in unstructured environments: How far have we come?. *arXiv preprint arXiv:2410.07701*.

Mohamed, M. F., Eltoukhy, M. M., Al Ruqeishi, K., & Salah, A. (2023). An adapted multi-objective genetic algorithm for healthcare supplier selection decision. *Mathematics*, *11*(6), 1537.

Oliveira, J. L. A., Aquino, A. L., Pinheiro, R. G., & Nogueira, B. (2024). Optimizing public transport system using biased random-key genetic algorithm. *Applied Soft Computing*, *158*, 111578.

Özer, M. M. (2024). Hybrid Genetic Algorithm Based on Machine Learning and Fitness Function Estimation Proposal for Ground Vehicle and Drone Cooperative Delivery Problem. In *Artificial Intelligence in Prescriptive Analytics: Innovations in Decision Analysis, Intelligent Optimization, and Data-Driven Decisions* (pp. 171-210). Cham: Springer Nature Switzerland.

Pan, Y., Yang, Y., & Li, W. (2021). A deep learning trained by genetic algorithm to improve the efficiency of path planning for data collection with multi-UAV. *Ieee Access*, *9*, 7994-8005.

Pandey, H. M., Chaudhary, A., & Mehrotra, D. (2014). A comparative review of approaches to prevent premature convergence in GA. *Applied Soft Computing*, *24*, 1047-1077.

Peffers, K., Tuunanen, T., & Niehaves, B. (2018). Design science research genres: introduction to the special issue on exemplars and criteria for applicable design science research.

Pétrowski, A., & Ben-Hamida, S. (2017). *Evolutionary algorithms* (Vol. 9). John Wiley & Sons.

Ramasamy, S., Reddinger, J. P. F., Dotterweich, J. M., Childers, M. A., & Bhounsule, P. A. (2022). Coordinated route planning of multiple fuel-constrained unmanned aerial systems with recharging on an unmanned ground vehicle for mission coverage. *Journal of Intelligent & Robotic Systems*, *106*(1), 30.

Rashid, A. B., Kausik, A. K., Al Hassan Sunny, A., & Bappy, M. H. (2023). Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International journal of intelligent systems*, *2023*(1), 8676366.

Ray, P. P. (2023). Benchmarking, ethical alignment, and evaluation framework for conversational AI: Advancing responsible development of ChatGPT. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *3*(3), 100136.

Sajadieh, S. M. M., & Noh, S. D. (2025). From Simulation to Autonomy: Reviews of the Integration of Artificial Intelligence and Digital Twins. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 1-32.

Saxon, D. (2021). *Fighting Machines: Autonomous Weapons and Human Dignity*. University of Pennsylvania Press.

Sayles, J. (2024). Aligning AI Governance with Other Internal Governance Models for Trustworthy AI:"The Convergence of Governance Frameworks". In *Principles of AI Governance and Model Risk Management: Master the Techniques for Ethical and Transparent AI Systems* (pp. 113-172). Berkeley, CA: Apress.

Sloss, A. N., & Gustafson, S. (2020). 2019 evolutionary algorithms review. *Genetic programming Theory and practice XVII*, 307-344.

Wu, Y., Li, X., Liu, Q., & Tong, G. (2022). The analysis of credit risks in agricultural supply chain finance assessment model based on genetic algorithm and backpropagation neural network. *Computational Economics*, *60*(4), 1269-1292.

Zhan, H., Zhang, Y., Huang, J., Song, Y., Xing, L., Wu, J., & Gao, Z. (2024). A reinforcement learning-based evolutionary algorithm for the unmanned aerial vehicles maritime search and rescue path planning problem considering multiple rescue centers. *Memetic Computing*, *16*(3), 373-386.

Zhang, F., Mei, Y., Nguyen, S., & Zhang, M. (2023). Survey on genetic programming and machine learning techniques for heuristic design in job shop scheduling. *IEEE Transactions on Evolutionary Computation*, *28*(1), 147-167.

Beyond Posters: A User-Centric Digital Twin Framework for Cybersecurity Awareness

Fhatuwani Makharamedzha, Errol Baloyi, Rendani Mmbodi and Ndabezinhle Hlongwane
Council for Scientific and Industrial Research (CSIR), Pretoria, South Africa

FMakharamedzha@csir.co.za
EBaloyi2@csir.co.za
RMmbodi@csir.co.za
NHlongwane@csir.co.za

Abstract: Traditional cybersecurity awareness (CSA) methods, such as posters, flyers, and static training modules often fail to engage users or drive lasting behavioural change. To address these limitations, this paper proposes a novel, user-centric approach to CSA using Digital Twin (DT) technology integrated with machine learning (ML). The proposed framework introduces the concept of a User-Centric Digital Twin (UCDT)-CSA, a dynamic digital replica of each user modelled on their cybersecurity knowledge, behaviours, and risk profile. While UCDTs have been applied in domains such as construction, aquaculture, and healthcare, this work pioneers their use in the cybersecurity context. The system begins with a pre-assessment to capture individual user responses, which are used to configure a personalized training path. Through ongoing interaction with adaptive simulations and scenario-based learning, the UCDT-CSA evolves in real time, enabling training that continuously adjusts to user performance and behaviour. ML models analyse these interactions to refine each twin's profile, delivering increasingly targeted content and interventions aimed at improving secure behaviours. This approach transforms CSA from a static, compliance-focused exercise into an engaging, data-driven, and behaviourally adaptive learning experience. The paper outlines the architecture of the UCDT-CSA framework, discusses key implementation considerations, and sets the stage for future empirical validation and deployment in government, Small and Medium-Sized Enterprises (SMEs) and academic environment.

Keywords: Digital twin, Machine learning, Cybersecurity awareness

1. Introduction

In today's technologically interconnected world, cybersecurity awareness (CSA) has become a critical component of online security. This change arises at a time when organizations are increasingly dependent on technology for communication, online transactions, and various operational activities. CSA is commonly defined as the knowledge, behaviours, and practices that enable individuals and organizations to identify threats and risks, and to adopt prudent measures within the digital environment (Maqousi, 2023). Nonetheless, CSA often fails to receive the level of attention it warrants, particularly among frequent Internet users who remain vulnerable to cyber threats and attacks due to insecure online practices. It is also important to emphasize that CSA is not solely the responsibility of Information Technology (IT) professionals, but a necessary competency for all organizational employees and individuals.

CSA involves everyone who uses technology, from employees managing sensitive data to individuals using their modern gadgets like smartphones in the cyber space (Maqousi, 2023). Thereby, making cybersecurity both a personal and collective responsibility. Human behaviour remains the biggest threat to cybersecurity. Past research shows that most frequent incidents have occurred because of human error and or negligence, which includes blindly clicking on malicious links, reusing past passwords, and a general lack of social engineering tactics (Tambe-Jagtap, 2023). In most cases, organizations do finance technical defences like encryption and detection systems which end up unsuccessful, as threat actors often exploit the human element, using tactics such as social-engineering and password theft.

Furthermore, modern threat actors are now conducting advanced modern attacks such as weaponizing artificial intelligence (AI), deepfakes and psychological manipulation to power up these scams, which further complicates the threat landscape (Mushtaq, 2019). The results of these attacks expand as far as financial losses and reputational damage. An attempt by some organization to curb the rise of cyber-threats through CSA training has been made, however, this includes CSA programmes delivered through traditional delivery methods such as posters, generic training modules, and compliance-driven campaigns which often fail to engage users effectively or adapt to their evolving risk profiles (Fujs, 2022).

These methods typically rely on one-size-fits-all content, ignoring the diversity in user behaviour, digital literacy, and security attitudes. To address this limitation, a need for personalization CSA is needed. As a such, the current study proposes a baseline user-centric digital twin (UCDT)-CSA framework. To provide a theoretical

foundation and the necessary background for the framework, existing literature on CSA and digital twin technology was reviewed. This involved searching search engines and academic databases such as ScienceDirect and Google Scholar for peer-reviewed publications and related reports.

2. Literature Review

To engineer significant personalization in CSA, innovative methods are needed. Digital Twin (DT) technology offers a novel approach. Digital Twins are dynamic virtual replicas of a physical system, objects that constantly grow in real time-based data from their original counterparts (Azambuja et al, 2024). They have recently been used in fields such as construction, healthcare, business and production (Alhamam, Rahman and Aljughaiman, 2025). As such, the current study intends to apply DTs in a human-centric manner, UCDT can serve as a living model of an individual's cybersecurity behaviors, risk profile and information. Below the current study investigates some of the CSA delivery methods.

2.1 Traditional Cybersecurity Awareness (CSA) Approaches

Traditional CSA often relies on mandatory annual training, posters, and emails. However, these kinds of methods are restricted and limited in impact and widely adopted in both educational and corporate environments (Alam, 2017). Whilst they may be easy to scale and cost-effective, prior studies have consistently demonstrated that their effectiveness in fostering sustainable security behaviour is limited. For instance, Parsons et al. (2017) realized that required training often results in short-term knowledge retention which also fails to create lasting behavioural change, an observation that was also noted by Alshaikh (2020).

2.2 Personalization and Adaptive Learning in CSA

The value of personalization is progressively recognized. For example, GPT-powered CSA has been shown to exceed static programs by dynamically adapting content to user profiles which also improve engagement (Al-Dhamari, 2024). Likewise, adaptive learning in cybersecurity training assists to deliver content suited to students' skill levels by avoiding unnecessary or overly complex material. Research shows that adaptive approaches improve engagement and completion rates, and students report better experiences (Seda, 2022).

2.3 Virtual Reality (VR) in Cybersecurity Awareness

Virtual Reality (VR) has lately been explored as one of the innovative medium for CSA together with training. Far from traditional or gamified approaches, VR gives out interactive environment that gives users a safely experience simulated cyber threats in real time (Sheridan et al, 2020). By creating realistic scenarios such as social engineering encounters, phishing attacks, or password exploitation, VR offers participants opportunity to learn through experiential engagement rather than passive instruction. This fascinating quality assist in overcoming the limitations of static awareness programs, which often fail to translate into real-world behaviour.

Research have also shown that VR-based cybersecurity training enhances both retention and engagement. For example, McGill and Klobas (2020) show that immersive training scenarios significantly enhance users' ability to detect phishing attempts com-pared to text-based learning. Comparably, Kritzinger and Solms (2021) argue that VR provides "situated learning," where students can develop skills within realistic contexts, leading to stronger transfer of knowledge into workplace practices. Further-more, VR can be customized to specific roles and environments, which allows personalized training experiences that account for individual differences in risk perception and cognitive style.

2.4 Digital Twin Technology in Cybersecurity and Beyond

DTs have obtained momentum across diverse domains such as healthcare, manufacturing, and human–machine collaboration, where they are used for predictive maintenance, patient monitoring, and real-time process optimization (Fuller et al, 2020; Tao et al, 2019; Minerva et al, 2020). Their core strength lies in creating dynamic virtual replicas of physical systems, enabling simulation, optimization, and monitoring without interfering with real-world operations.

Even though initially applied in industrial and engineering contexts, researchers have increasingly explored DTs within cybersecurity. For example, DTs have been employed for threat modelling, intrusion detection, and resilience testing in Industrial Control Systems (ICS) and IoT infrastructures (Cimino et al, 2019; Grieves and Vickers, 2017). They allow corporations to simulate red- and blue-team exercises, test incident response strategies, and rehearse chaos scenarios without disrupting production systems. Other works show how DTs

can aid in anomaly detection, forensic analysis, and security control validation by enabling replay and traceability of cyber events (Jones et al, 2020; Yaqoob et al, 2020).

Nonetheless with these promising applications, the user-centric potential of DTs for CSA and education remains underexplored. While DTs have been leveraged to protect infrastructure and validate technical defences, few if any frameworks explicitly focus on enhancing end-user awareness, decision-making, and behavioural adaptation in the face of cyber threats. To the best of the authors' knowledge, no prior work has proposed a digital twin–driven model specifically designed for cybersecurity training and awareness, highlighting a critical gap that this study aims to address.

2.5 Gaps and Opportunities

While DTs are deep-rooted for systems and network modelling, their use in human-centric CSA, mainly at the isolated level, remains evolving. Existing DT studies focuses on infrastructure, enterprise networks, or cyber-physical systems, not only on modelling user behaviour, knowledge, or learning trajectories (Ghosh, 2024). There's an important opportunity to fill this opportunity by developing UCDTs which is a virtual representation of user cognition, habits, and progression. Similar with UCDTs could enable personal scenario simulation, adaptive feedback, and training personalization driven by ML. This approach shifts CSA from one-size-fits-all modules to one of dynamic, behavioural, and user-specific learning paths bridging current CSA limitations, adaptive learning, and DT capabilities.

3. Conceptual Framework Design: UCDT-CSA

This section outlines the UCDT-CSA framework, as noted, to the best of the author's knowledge, this is the first framework to apply a UCDT approach specifically to CSA. The proposed UCDT-CSA framework is designed to model and influence individual user behaviour through a modular, ML-enhanced CSA system. This section describes the high-level architecture, data flow, and core functional components of the framework. Figure 1 provides an illustration of the UCDT-CSA framework.

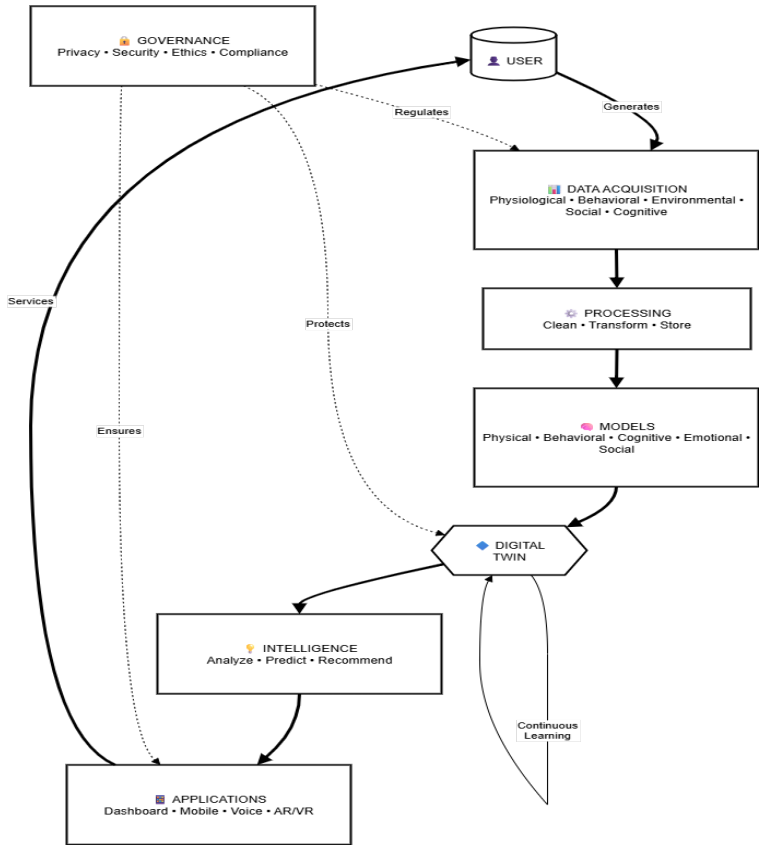


Figure 1: UCDT-CSA Framework Design

As illustrated in Figure 1, the UCDT-CSA framework is designed around a modular, engine-based architecture comprising three layers: data acquisition, processing, and models. Each layer is responsible for specific tasks, enabling independent development, seamless integration, and future scalability. The data acquisition layer

captures initial and ongoing user data to build detailed user profiles. The processing layer cleans, transforms, and stores these profiles, while the model layer applies AI/ML techniques to personalize cybersecurity training and maintain the DT engine through continuous learning. Finally, the framework includes an application layer that delivers adaptive cybersecurity awareness content.

4. Implementation Considerations

The successful deployment of UCDT-CSA framework requires careful attention to multiple implementation dimensions that extend beyond technical architecture. This section examines four critical considerations that organizations must be addressed to ensure effective, ethical, and sustainable UCDT-CSA framework implementation.

5. Privacy and Ethical Concerns

The implementation of user-centric digital twins raises fundamental privacy and ethical challenges that must be addressed proactively. Data sovereignty represents a primary concern, as users must maintain control over their personal information while enabling the system to function effectively (Minerva et al., 2020). The principle of informed consent becomes complex when dealing with continuous data collection and evolving model capabilities, requiring dynamic consent mechanisms that adapt to changing data uses (Tao et al, 2019). Algorithmic transparency poses another critical challenge. Users have the right to understand how their digital twin makes predictions and recommendations, necessitating explainable AI approaches that balance model complexity with interpretability (Fuller et al, 2020). The implementation must address data minimization principles, collecting only necessary information while maintaining model effectiveness (El Saddik, 2018).

Organizations must implement differential privacy techniques to protect individual privacy while enabling population-level insights (Qi et al, 2021). This includes implementing privacy-preserving federated learning approaches that allow model training without centralizing sensitive data (Zhang et al, 2022). The framework must also address algorithmic bias through continuous monitoring and correction mechanisms, ensuring equitable outcomes across diverse user populations (Barricelli et al, 2019). Ethical AI governance structures must be embedded within the implementation, including ethics review boards, regular audits, and clear accountability frameworks (van der Valk et al, 2021). The system should implement right to explanation capabilities, allowing users to understand and challenge automated decisions that significantly affect them (Liu et al, 2021).

5.1 System Integration

Successful digital twin implementation requires seamless integration with existing organizational infrastructure and third-party systems. Legacy system compatibility presents a significant challenge, as organizations must bridge modern digital twin architectures with established enterprise systems (Kritzinger et al., 2018). This requires developing robust middleware solutions and API standardization to ensure interoperability across heterogeneous technology stacks (Tao et al, 2018). Data integration complexities arise from the need to harmonize information from multiple sources with varying formats, quality levels, and update frequencies (Zheng et al, 2021). Organizations must implement master data management strategies and data governance frameworks to ensure consistency and reliability across integrated systems (Semeraro et al, 2021).

The implementation must address real-time synchronization requirements, ensuring that the digital twin maintains temporal coherence with physical world changes (Jones et al, 2020). This necessitates event-driven architectures and stream processing capabilities that can handle high-velocity data flows while maintaining system stability (Rasheed et al, 2020). Semantic interoperability represents another critical challenge, requiring the development of common ontologies and data models that enable meaningful information exchange between systems (Jacoby and Usländer, 2020). Organizations must implement service-oriented architectures that support loose coupling and enable gradual migration from legacy systems (Singh et al, 2021).

Security integration across systems requires implementing consistent authentication and authorization mechanisms while maintaining zero-trust principles (Alcaraz and Lopez, 2022). This includes establishing secure API gateways, identity federation, and encrypted communication channels across all integrated components (Dietz and Pernul, 2020).

5.2 Scalability

Digital twin implementations must be designed to scale efficiently as user bases grow, and data volumes expand exponentially. Horizontal scalability requires architectures that can distribute processing across multiple nodes without degrading performance (VanDerHorn and Mahadevan, 2021). This necessitates implementing microservices architectures with container orchestration platforms that enable dynamic resource allocation (Minerva et al, 2020). Data scalability challenges emerge as systems must process and store massive volumes of heterogeneous data while maintaining query performance (Lehner and Sattler, 2020). Organizations must implement distributed storage solutions, data partitioning strategies, and intelligent caching mechanisms to manage growing data repositories effectively (Qi and Tao, 2018).

Model scalability becomes critical as the number of concurrent digital twins increases, requiring efficient model serving infrastructure that can handle thousands of simultaneous predictions (Lim et al, 2021). This includes implementing model compression techniques, edge computing strategies, and federated learning approaches that distribute computational load (Park et al, 2022). Network scalability must accommodate increasing data transmission requirements while maintaining low latency for real-time applications (Lu et al, 2020). Organizations should implement content delivery networks, edge caching, and 5G integration to ensure responsive system performance at scale (Minerva et al, 2021).

Cost scalability represents a crucial consideration, as infrastructure costs can grow non-linearly with system expansion (Cimino et al, 2019). Implementation strategies must include cloud-native architectures, serverless computing, and auto-scaling policies that optimize resource utilization while controlling expenses (Alam and El Saddik, 2017).

5.3 Content Development

The creation and maintenance of digital twin content require sophisticated development frameworks and continuous refinement processes. Model development pipelines must support iterative refinement, version control, and A/B testing capabilities to ensure model quality and relevance (Zheng et al, 2022). This includes establishing MLOps practices that standardize model development, testing, and deployment workflows (Rathore et al, 2021).

Personalization algorithms must balance individual customization with computational efficiency, requiring adaptive learning mechanisms that evolve with user behaviour (Wang et al., 2022). Organizations must implement transfer learning and meta-learning approaches that enable rapid personalization for new users while maintaining model accuracy (Lv et al., 2022). Content validation processes must ensure that digital twin outputs remain accurate, relevant, and beneficial to users (Boje et al., 2020). This requires implementing continuous monitoring systems, feedback loops, and quality assurance frameworks that detect and correct model drift (Errandonea et al, 2020).

Knowledge representation challenges arise in encoding complex human behaviours and contexts within computational models (Leng et a., 2021). Organizations must develop multi-modal representation learning approaches that capture the richness of human experience while maintaining computational tractability (Kang et al, 2021). User experience design for digital twin interfaces requires careful attention to information presentation, interaction paradigms, and cognitive load management (Kaur et al., 2020). Implementation must include adaptive interfaces, progressive disclosure patterns, and context-aware visualizations that make complex information accessible to diverse user populations (Madni et al, 2019).

5.4 Implementation Framework Integration

These implementation considerations must be integrated holistically within the overall digital twin architecture. The following framework illustrates how these considerations interact with the technical layers:

Privacy-Preserving Architecture

- Implement homomorphic encryption for sensitive data processing.
- Deploy differential privacy mechanisms at data aggregation points.
- Establish consent management systems with granular control options.
- Enable data portability and deletion capabilities per GDPR requirements.

Integration Architecture

- Deploy enterprise service bus (ESB) for system communication.

- Implement API management platforms with versioning support.
- Establish data transformation services for format harmonization.
- Create adapter patterns for legacy system connectivity.

Scalable Infrastructure

- Utilize Kubernetes for container orchestration and auto-scaling.
- Implement Apache Kafka for distributed event streaming.
- Deploy distributed databases with sharding capabilities.
- Establish CDN and edge computing nodes for global reach.

Content Management System

- Create model registries for version control and governance.
- Implement feature stores for consistent feature engineering.
- Establish A/B testing frameworks for model evaluation.
- Deploy continuous integration/continuous deployment (CI/CD) pipelines.

6. Use Cases

The proposed UCDT-CSA framework offers versatile applications in CSA, addressing diverse needs across various organizational environments. The key potential use cases outlined below demonstrate its practical value. It should be noted that while this framework is primarily intended for adoption by organizations in South Africa, its applicability is not limited to this context and can be extended globally. However, the focus of these use cases remains on the South African environment.

6.1 Government Agencies

Government agencies are responsible for safeguarding sensitive information and delivering public services efficiently, involving employees at all levels from general staff to technical personnel, all of whom need to possess a basic understanding of CSA to ensure effective service delivery (Ngoma, Keevy and Rama, 2021). Government departments are frequent targets of cyberattacks. In South Africa, Siphambili et al (2024) noted that since 2019, the government sector has been predominantly targeted by ransomware attacks, with ten different departments affected and one department attacked twice. The study recommended implementing a comprehensive CSA programme as a first line of defence against such threats. This is where the UCDT-CSA framework becomes relevant, as it is well-suited for this environment and can tailor CSA content to meet the needs of different user groups.

6.2 Educational Institutions

Tertiary institutions and schools often deliver CSA to a diverse user base with varying technical backgrounds. The UCDT-CSA framework is well-suited for these environments, as it can differentiate content for students, faculty, and administrative staff. If adopted, it could also increase engagement through interactive content, which, according to Nkongolo (2024), is often preferred by this demographic. Furthermore, it can be used as a standardized framework across different Higher Education Institutions (HEIs), which could make a significant difference. Masenya's (2023) study, which investigated the CSA of students within South African HEIs, where this framework will also probable be implemented found that while students are aware of some cybersecurity terms, there is a lack of collaboration within these institutions on the best approach to CSA training.

6.3 Small and Medium-Sized Enterprises (SMEs)

A study by Eybers and Mvundla (2022) investigated CSA among managers in fifteen small and SMEs in South Africa. The findings revealed that most SME managers are aware of cybersecurity threats, as they are currently facing them within their organizations. However, the challenge lies with employees, many of whom lack such awareness. Furthermore, SMEs often lack the financial resources to provide comprehensive cybersecurity training programs, and many outsource their IT and security functions. In this context, the framework is particularly valuable, as it enables the customization of CSA content to address the specific needs of different roles within an organization, such as management, technical staff, and general employees. The framework can therefore offer a scalable and cost-effective solution for enhancing cybersecurity awareness across various organizational levels.

7. Conclusion and Future Work

This paper addresses the identified gap by proposing a baseline UCDT-CSA framework that builds a dynamic digital profile of the user through initial assessment and continuous behavioural input, integrates AI/ML components to personalize awareness delivery over time, offers a scalable and modular approach for use across organizational settings and positions the user as an active participant in the cybersecurity lifecycle, not just a passive recipient of awareness content. By bridging concepts from DT technology, adaptive learning, and cybersecurity behaviour modelling, this work introduces a novel paradigm for CSA, one that is personalized, data-driven, and continuous.

For future work, a working prototype of the UCDT framework incorporating the core engines, Pre-Assessment, Digital Twin Management, AI/ML Adaptation, and Awareness Delivery will be developed. This will enable practical experimentation and refinement of the system components. This framework will be deployed within diverse organizational settings to assess its effectiveness in improving cybersecurity knowledge and behaviour.

Ethics Declaration: Ethical clearance was not obtained, since the current study did not involve human participants.

AI Declaration: AI tools were not used in the current study.

References

Alam, M., and El Saddik, A. (2017) "A digital twin architecture reference model for the cloud-based cyber-physical systems" *IEEE Access*, Vol. 5, 2050-2062.

Al-Dhamari, R. (2024). AI-driven adaptive cybersecurity awareness using GPT models: A user-focused approach. Computers & Security, 141, 103365. https://doi.org/10.1016/j.cose.2024.103365

Alcaraz, C., and Lopez, J. (2022) "Digital twin: A comprehensive survey of security threats", *IEEE Communications Surveys & Tutorials*, Vol. 24, No. 3, pp 1475-1503.

Alhamam, N., Rahman, M. M. H. and Aljughaiman, A. (2025) "A comprehensive review on cybersecurity of digital twins: Issues, challenges, and future research directions", *IEEE Access*, Vol. 13, pp 45106–45124.

Azambuja, A. J. G. de, Giese, T., Schützer, K., Anderl, R., Schleich, B., and Almeida, V. R. (2024) "Digital twins in Industry 4.0 – Opportunities and challenges related to cyber security", *Procedia CIRP*, pp 25–30.

Alshaikh, M. (2020). Developing cybersecurity culture to influence employee behavior: A practice perspective. Computers & Security, 98, 102003. https://doi.org/10.1016/j.cose.2020.102003

Barricelli, B. R., Casiraghi, E., and Fogli, D. (2019) "A survey on digital twin: Definitions, characteristics, applications, and design implications" *IEEE Access,* 7, 167653-167671.

Boje, C., Guerriero, A., Kubicki, S., and Rezgui, Y. (2020) "Towards a semantic Construction Digital Twin: Directions for future research" *Automation in Construction*.

Cimino, C., Negri, E. and Fumagalli, L. (2019) "Review of digital twin applications in manufacturing", *Computers in Industry*.

Dietz, M., and Pernul, G. (2020) "Digital twin: Empowering enterprises towards a system-of-systems approach" *Business & Information Systems Engineering*, Vol. 62, Vol. 2, pp 179-184.

El Saddik, A. (2018) "Digital twins: The convergence of multimedia technologies", *IEEE Multimedia*, Vol. 25, No. 2, pp 87-92.

Errandonea, I., Beltrán, S., & Arrizabalaga, S. (2020) "Digital twin for maintenance: Data-driven approach", *Procedia Manufacturing*, Vol. 42, pp 362–369.

Eybers, S. and Mvundla, Z. (2022) "Investigating Cyber Security Awareness (CSA) Amongst Managers in Small and Medium Enterprises (SMEs)" *Comprehensible Science. ICCS 2021. Lecture Notes in Networks and Systems*, Vol. 315.

Fujs, D., Vrhovec, S. and Vavpotič, D. (2022) "Towards personalized user training for secure use of information systems", *The International Arab Journal of Information Technology*, Vol. 19, No. 3.

Fuller, A., Fan, Z., Day, C., and Barlow, C. (2020) "Digital twin: Enabling technologies, challenges and open research" *IEEE Access*, Vol. 8, 108952-108971.

Jacoby, M. and Usländer, T. (2020) "Digital twin and internet of things—Current standards landscape", *Applied Sciences*, Vol. 10, No. 18, 6519.

Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unanticipated failures in complex systems. In Proceedings of the ASME 2017.

Ghosh, S. (2024). User-centric digital twins for adaptive cybersecurity awareness: A conceptual framework. IEEE Internet of Things Journal, 11(2), 1454–1466. https://doi.org/10.1109/JIOT.2023.3312345.

Jones, D., Snider, C., Nassehi, A., Yon, J. and Hicks, B. (2020) "Characterising the Digital Twin: A systematic literature review", *CIRP Journal of Manufacturing Science and Technology,* 29, pp 36-52.

Kang, J., Wang, D., and Chen, Y. (2021) "Multimodal digital twin: A conceptual framework for human-centric applications", *Journal of Manufacturing Systems*, 61, pp 424-435.

Kaur, S., Singh, A., & Kumar, P. (2020) "User-centric visualization for digital twins" *International Journal of Advanced Computer Science and Applications*, Vol. 11, No. 5, pp 233–241.

Kritzinger, W., Karner, M., Traar, G., Henjes, J. and Sihn, W. (2018) "Digital Twin in manufacturing: A categorical literature review and classification", *IFAC-PapersOnLine,* Vol. 51, No. 11, 1016-1022.

Lehner, W. and Sattler, K. U. (2020) "Database support for data-driven digital twins", *Datenbank-Spektrum*, Vol. 20, No. 2, pp 119-126.

Leng, J., Wang, D., Shen, W., Li, X., Liu, Q. and Chen, X. (2021) "Digital twins-based smart manufacturing system design in Industry 4.0: A review" *Journal of Manufacturing Systems*, 60, 119-137.

Lim, K. Y. H., Zheng, P., and Chen, C. H. (2021) "A state-of-the-art survey of Digital Twin: Techniques, engineering product lifecycle management and business innovation perspectives" *Journal of Intelligent Manufacturing*, Vol .31, No. 6, 1313-1337.

Liu, M., Fang, S., Dong, H. and Xu, C. (2021) "Review of digital twin about concepts, technologies, and industrial applications", *Journal of Manufacturing Systems*, 58, 346-361.

Lu, Y., Liu, C., Kevin, I., Wang, K., Huang, H. and Xu, X. (2020) "Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues", *Robotics and Computer-Integrated Manufacturing*, 61, 101837.

Lv, Z., Xie, S., Li, Y., Hossain, M. S. and El Saddik, A. (2022) "Building the metaverse by digital twins at all scales, state, relation" *Virtual Reality & Intelligent Hardware*, Vol. 4, No. 6, 459-470.

Madni, A. M., Madni, C. C., and Lucero, S. D. (2019) "Leveraging digital twin technology in model-based systems engineering" *Systems*, Vol. 7, No. 1, 7.

Maqousi, A. A. (2023) "A proposed framework for user cybersecurity awareness" *24th International Arab Conference on Information Technology (ACIT)*, 1–6.

Masenya, T. M. (2023) "Awareness and knowledge of cyber ethical behaviour by students in higher education institutions in south africa", *Handbook of Research on Cybersecurity Risk in Contemporary Business Systems*, pp 33–48.

McGill, T. J., & Klobas, J. E. (2020). Immersive cybersecurity training using VR: Improving phishing detection skills. Behaviour & Information Technology, 39(12), 1343–1358. https://doi.org/10.1080/0144929X.2019.1699691.

Minerva, R., Crespi, N., Awan, F. M. and Rotondi, D. (2021). Digital Twins: Properties, Software Frameworks, and Application Scenarios. *IT Professional*, Vol. 23, No. 1, pp 51-56.

Minerva, R., Lee, G. M., and Crespi, N. (2020) "Digital twin in the IoT context: A survey on technical features, scenarios, and architectural model" *Proceedings of the IEEE*, Vol. 108, No. 10, pp 1785-1824.

Mushtaq, S. (2019) "Modern cyber-attacks and cloud security: Strengthening information security in emerging technologies".

Ngoma, M. L., Keevy, M., and Rama, P. (2021) "Cyber-security awareness of South African state-mandated public sector organisations" *Southern African Journal of Accountability and Auditing Research*, Vol. 23, No. 1, pp 53–64.

Nkongolo, M. (2024). CyberMoraba: A game-based approach enhancing cybersecurity awareness. International Conference on Cyber Warfare and Security, 19(1), 240–250. https://doi.org/10.34190/iccws.19.1.1957

Parsons, K., McCormac, A., Butavicius, M., Pattinson, M., & Jerram, C. (2017). Determining employee cybersecurity behavior: The role of awareness training. Computers & Security, 68, 24–34. https://doi.org/10.1016/j.cose.2017.04.005

Park, K. T., Nam, Y. W., Lee, H. S., Im, S. J., Noh, S. D., Son, J. Y. and Kim, H. (2022) "Design and implementation of a digital twin application for a connected micro smart factory*", International Journal of Computer Integrated Manufacturing*, Vol. 32, No. 6, pp 596-614.

Qi, Q. and Tao, F. (2018) "Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison", *IEEE Access*, 6, 3585-3593.

Qi, Q., Tao, F., Hu, T., Li, F. and Cheng, J. (2021) "Enabling technologies and tools for digital twin" *Journal of Manufacturing Systems*, 58, pp 3–21.

Rasheed, A., San, O. and Kvamsdal, T. (2020) "Digital twin: Values, challenges and enablers from a modeling perspective" *IEEE Access*, 8, 21980-22012.

Rathore, M. M., Shah, S. A., Shukla, D., Bentafat, E. and Bakiras, S. (2021) "The role of AI, machine learning, and big data in digital twinning: A systematic literature review, challenges, and opportunities" *IEEE Access*, 9, 32030-32052.

Seda, P. (2022). Adaptive learning for cybersecurity awareness in higher education. Education and Information Technologies, 27(8), 10721–10739. https://doi.org/10.1007/s10639-022-11035-2

Semeraro, C., Lezoche, M., Panetto, H. and Dassisti, M. (2021) "Digital twin paradigm: A systematic literature review", *Computers in Industry,* 130, 103469.

Singh, M., Fuenmayor, E., Hinchy, E. P., Qiao, Y., Murray, N. and Devine, D. (2021) "Digital twin: Origin to future" *Applied System Innovation*, Vol. 4, No. 2, 36.

Siphambili, N., Mahlasela, O., Baloyi, E., & Mukondeleli, E. (2024) "A Review of the South African Public Sector's Capability in Combating Ransomware", *4th International Multidisciplinary Information Technology and Engineering Conference (IMITEC),* pp 493-499.

Sheridan, T. B., Williams, C., & Farooq, S. (2020). Immersive VR for cyber threat education. Computers & Security, 94, 101860. https://doi.org/10.1016/j.cose.2020.101860.

Tambe-Jagtap, S. N. (2023) "Human-centric cybersecurity: Understanding and mitigating the role of human error in cyber incidents" *SHIFRA*, 53–59.

Tao, F., Cheng, J., Qi, Q., Zhang, M., Zhang, H., and Sui, F. (2018) "Digital twin-driven product design, manufacturing and service with big data" *The International Journal of Advanced Manufacturing Technology*, 94(9), 3563-3576.

Tao, F., Zhang, H., Liu, A. and Nee, A. Y. (2019) "Digital twin in industry: State-of-the-art" *IEEE Transactions on Industrial Informatics*, Vol. 15, No. 4, 2405-2415.

van der Valk, H., Haße, H. and Breitner, M. H. (2021) "Ethical governance for AI-driven digital twins", *AI and Ethics*, 1, pp 303–315.

VanDerHorn, E., & Mahadevan, S. (2021) "Digital Twin: Generalization, characterization and implementation" *Decision Support Systems*, 145, 113524.

Wang, B., Zhou, H., Li, X., Yang, G., Zheng, P., Song, C., ... and Wang, L. (2022) "Human Digital Twin in the context of Industry 5.0" *Robotics and Computer-Integrated Manufacturing,* 85, 102626.

Zhang, H., Liu, Y. and Zhou, D. (2022) "Privacy-preserving federated learning for digital twins*", IEEE Internet of Things Journal*, Vol. 9, No. 10.

Zheng, P., Lin, T. J., Chen, C. H., & Xu, X. (2021). A systematic design approach for service innovation of digital twin-based services. European Journal of Operational Research, 295(2), 646-661.

Zheng, Y., Yang, S., and Cheng, H. (2022) "An application framework of digital twin and its case study", *Journal of Ambient Intelligence and Humanized Computing*, Vol. 10, No. 3, 1141-1153.

Architectural Framework for an Enhanced Multi-Party Fully Homomorphic Encryption Scheme

Joshua Mamza[1], Idris Ismaila[1], Joseph Ojeniyi[1], Shafi'i Abdulhamid[2], Moses Noel[1] and Olusanjo Fasola[1]

[1]Department of Cyber Security, Federal University of Technology Minna, Niger State, Nigeria
[2]Department of Information Technology, Community College of Qatar, Qatar

Joshua.pg6759@st.futminna.edu.ng
Ismi.idris@futminna.edu.ng
ojeniyija@futminna.edu.ng
shafii.abdulhamid@ccq.edu.qa
moses.noel@futminna.edu.ng
sanjo.fasola@futminna.edu.ng

Abstract: The Common Vulnerability Scoring System (CVSS) depends on reliable vulnerability data from expert, but the current process of vulnerability score generation and transmission remain exposed to data manipulation and interception. Existing research work used supervised machine learning to automate CVSS scoring with up to 90% accuracy, but their plaintext-based approach lacked cryptographic protections, leaving it vulnerable to Man-in-the-Middle (MitM) attacks. Another research work introduced a homomorphic encryption-based framework that preserves data confidentiality during computation and offers moderate performance gains. However, their dependance on a single trusted aggregator, static key management, and absence of dynamic integrity threshold mechanisms left the system exposed if the aggregator's key or channel were compromised. An architectural framework for an Enhanced Multi-Party Fully Homomorphic Encryption Scheme (EMHES) was designed to combat Man-in-the-Middle (MitM) attacks targeting Vulnerability Score manipulation. By employing Homomorphic Encryption, the framework enables computations on encrypted vulnerability scores, ensuring confidentiality throughout their lifecycle. Key enhancements include integrating digital signatures to authenticate classified scores before encrypted transmission to cloud environments and verify the integrity of decrypted results post-processing. Digital signatures and regulatory oversight significantly strengthen security properties like non-repudiation, integrity, and confidentiality for cloud-based data computations. The EMHES architecture features a secure transmission channel with multiple security layers within the cloud service provider infrastructure. Additional security mechanisms include secure key management protocols, zero-knowledge proofs for integrity verification, and a resilient secure aggregation protocol designed to counter MitM attacks. From a computational analysis, baseline algorithms exhibit constant time complexity O(1), while the EMHES architecture operates with linear time complexity O(n). The result shows that EMHES provides superior security, integrity and performance on large datasets.

Keywords: Secure multiparty, Homomorphic encryption, Man-in-the-Middle attacks, Vulnerability score manipulation, EMHES

1. Introduction

The increasing interconnectivity of systems in serval domains that include healthcare, finance and cyber security, has highlighted the need of a secured data processing and sharing in terms of collaboration. Vulnerability Score (VS) plays a vital role in accessing the security position of these systems, organizations are guided in risk management approaches. Nonetheless, the manipulation of these scores poses significant threats through Man-in-the-Middle (MitM) attacks (Adablanu et al. 2024). VS systems such as the Common Vulnerability Scoring System (CVSS) are tools used for assessing and prioritizing security vulnerabilities using standard metrics used to evaluate the severity of vulnerability and potential impact. MitM attack significantly pose a threat by allowing malicious actors to intercept and alter vulnerability scores (Alanazi, Mahmood & Chowdhury 2023). Traditional encryption while deploying client confidentiality ensures decryption before computations are performed which creates vulnerability. Homomorphic encryption is a specialized encryption enables direct computations on an encrypted VS. Mathematical operations such as addition, multiplication is performed on encrypted file(Adablanu et al. 2024; Chen et al. 2024; Kamble, Jiet & Puri 2024).This paper proposes an architectural design for an Enhanced Multi-Party Fully Homomorphic Encryption Scheme (EMHES) to mitigate MitM attack against VS manipulation.

2. Background

2.1 Secure Multi-Party Computation (SMPC)

Secure Multi-party Computations (SMPC) is one of the important pillars in modern cryptography. Conversely, an Multi-party Computations (MPC) protocol allows a set of mutual distrust parties' i.e

$$P = (P_1, P_2, \ldots P_n) \quad (1)$$

To securely perform any computation or operation over their private input without revealing anything additional about their inputs. In most MPC protocols, the distrust is modelled by a centralized adversary A, who can corrupt and control parties during protocol execution. In perfect security, where A is a computationally unbounded adversary who can enforce the corrupt parties to behave arbitrary during protocol execution and where all the security guarantee are achieve without any error. Usually, the corruption capacity of A is modelled through a public known threshold *t.* where this is assumed. A can corrupt any parties up-to *t.*(Appan, Chandramouli & Choudhury 2023; Appan, Chandramouli & Choudhury 2025; Patil & Patra 2025). Private-preserving language model based on a SMPC addresses the privacy concerns associated with using large scale language model in Machine Learning as a Service (MLaaS). A scheme built on SMPC technology was propose to mitigate privacy leaking risks when processing sensitive data of a user. The proposed solution involves three non-colluding parties namely the data provider the model provider and the computing power provider. Compared to direct inference on large pre-trained models, this framework significantly improved inference speed by 1.55 to 6.25 times. The proposed solution goals stressed correctness privacy verifiability and efficiency (Song, Huang & Hu 2024; Yan et al. 2025; Zeng et al. 2025). SMPC are concerns with protocol execution coming under attack by an adversary which may corrupt one or more of the users to learn private information or cause the result said computation not to be correct. MPC are designed to prevent such kind of attack from being successful and can be proven mathematically to guarantee confidentiality, integrity and correctness (Pentyala et al. 2022; Yuan et al. 2021) . A secure and private multi-party deep learning through Differential Private – SMPC protocol which addresses security concerns about data privacy when it is train thereby achieving important improvement in communication and efficiency and speed compared to existing protocol (Das et al. 2025). SMPC perform an important role in block chain by allowing different parties to collaborate and compute on encrypted data without revealing their respective sensitive information. This process enhances privacy protection and data security block chain networks, it enables the block chain applications to execute complex task and smart contracts while safeguarding user privacy. SMPC delivers a secures and flexible data processing technique for block chain. It is used in the areas of finance, healthcare and supply chain (Bao et al. 2024; Ghanem & Moursy 2019; Zhou et al. 2021).

2.2 Vulnerability Scoring Manipulation

The management of vulnerabilities, which is a critical component of risk and resiliency efforts, typically is an ongoing process that involves the identification, classification, prioritization, and possibly remedy the vulnerabilities in devices likely to be targeted by adversaries to compromise network architecture and its components. Effective and efficient vulnerability management necessitates the allocation of resources such as time, personnel, and financial investment; thus, vulnerabilities should be prioritized according to their threat level (Jiang et al. 2025; Keskin et al. 2021; da Ponte, Rodrigues & Mattos 2023). Common Vulnerability Scoring System (CVSS) serves as a framework for assessing and communicating the severity of vulnerabilities and their potential impacts on devices. It provides a score that strongly echoes the severity of the vulnerability, enabling organizations and users to prioritize their remediall actions. The CVSS score is derived from metrics that evaluate various aspects of a vulnerability, precisely its impact on the confidentiality, integrity, and availability of the network system. (Tom & Kosacki 2023). A comparative analysis of large language models (LLMs) GPT and BERT for automated vulnerability scoring in cybersecurity field. It highlights that BERT outshines in understanding contextual relationships through bidirectional processing. GPT on the other hand, demonstrates strong generative capabilities but is more resource-intensive. The models' performance was evaluated on vulnerability classification tasks using the National Vulnerability Database (NVD) and finds that MediumBERT and SmallBERT outperform larger models like BERT and DistilBERT. In-addition, fine-tuning Generative Pre-trained Transformer (GPT) models significantly enhances their classification accuracy, particularly in specific vulnerability categories (Hashemi Chaleshtori & Ray 2023; Kalouptsoglou et al. 2024; Mirtaheri et al. 2025)

3. Related Work

(Sanon et al.2023 .) integrates federated learning (FL) with homomorphic encryption for predicting network traffic. This approach enhances privacy-preserving data analysis within a simulated environment of 5G. While the Cheon-Kim-Kim-Song (CKKS) encryption scheme requires increased computation time and inherent precision loss, the secure federated learning method demonstrates superior performance compared to traditional techniques. The results show that, despite the longer processing time required for the secure methods, ongoing advancements in homomorphic encryption are anticipated to improve efficiency in future applications.

A Multiparty Quantum Homomorphic Encryption (MQHE) scheme that allows multiple party perform quantum computations on their private data without requiring decryption, even in the faced with a dishonest server. The technique leverages measurement device independent quantum key Measurement-Divide-Independence Quantum Key Distribution (MDI-QKD) for secure generation and works with non-minimal states for error correction in computations. The proposed scheme outlines significant advantages in data protection during quantum computations, showing the application of quantum homomorphic encryption in secure multi party computation situations. Nonetheless, its complexity of implementation, potential computational overhead and assumptions about the behavior the server, limited scope of quantum operations present challenges that has to be addressed for deployment in several applications practically (Zhang et al., 2022; Li et al. 2025; Zhang et al. 2021). SMPCs perform an important role in block chain by allowing different parties to collaborate and compute on encrypted data without revealing their respective sensitive information. This process enhances privacy protection and data security block chain networks, it enables the block chain applications to execute complex task and smart contracts while safeguarding user privacy. SMPC delivers a secures and flexible data processing technique for block chain. It is used in the areas of finance, healthcare and supply chain (Bao et al. 2024) . Assessment testbed for cyber vulnerabilities by considering their result on business processes of an organization, beyond the standard CVSS metrics. It shows that Vulnerabilities with high CVSS scores do not always correlate with significant business impacts, urging decision-makers to prioritize vulnerabilities based on their true effects on organizational operations (Keskin et al. 2021). A comparative analysis of several FL techniques that emphasized their application in network traffic prediction. It discusses the significance of FL in data preservation privacy although collaborative model training among multiple organization. Accuracy and robustness are effectively identified, especially in the presence of attackers. The research focuses on application of homomorphic encryption alongside median aggregation. Additionally, the research also addresses the existing gaps regarding the comparative effectiveness of FL techniques creating way for improvement in the techniques of network traffic analysis (Sanon et al. 2024).

4. Research Methodology

The Enhanced Multi-Party Homomorphic Encryption Scheme (EMHES) has an implementation process that are structured and in multi-stage process comprising three main procedures: EMHES implementation stages, the time complexity analysis, and the space complexity analysis. This comprehensive approach ensures full functionality and performance of the designed EMHES.

4.1 EMHES Implementation Steps

1. Define, identification and retrieval of vulnerability score of

Retrieve Vulnerability Scores is equal to

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots x_{2n} \\ x_{31} & x_{32} & x_{33} & \dots x_{3n} \\ x_{m1} & x_{m2} & x_{m3} & \dots\ x_{mn} \end{bmatrix}, \qquad (2)$$

n = 12, m=104,976

2. An Architecture and flow chat have been developed to conceptualize the EMHES. Also, pseudocodes were been documents and analyzed to determine the efficiency of EMHES.

4.1.1 Time complexity analysis

The following steps outline the process for computing and analyzing the time complexity of the A = $\{A_i\ A_j A_k\}$ where the three key components are $A_i, A_j\ and\ A_k$ called Algorithm line number, Algorithm statement and Frequency count respectively. Where A equals sets of key components identified. A_i = 1, 2…n. Where $A_j =$ Statement 1, Statement 2…, Statement which represents the computations performed at each line of algorithm.

A_k = Count1, Count2 …, Count*n* which is computations performed at each Algorithm line. Where Count*i* is the number of times each operation is executed A = $\{A_i, A_j, A_k\}$. Also, let *f* frequency count of a computation. *f = 0 defines no computations performed, f = 1, defines computation is executed ones f= 1+1, defined computations is performed twice*, *f = n, n+1, defines a critical count in the loop during key generation and distribution*. The cumulative of the count F is represented as $\sum_{i=1}^{A_k}\{fA_k\}$. $A_k is\ total\ number\ of\ computations$, $fA_k\ is\ the\ frequency\ count\ in\ k-th\ computations$. Finally, *T* (n) = 6n +*x.*

where *T* is the algorithm time complexity which increases with the number of participants *n and x* is the constant.

4.1.2 Space complexity analysis

Sc is Space Complexity for Code per Character, fSccum is Cumulative frequency for code/Character, S_V is Space Complexity for Variable used in the algorithmic statement fSvcum is Cumulative count for space complexity for variable, S_{DC} is Space Complexity for Data Structure in the algorithmic statement

fS_{DC}cum is Cumulative count for space for data structure, S_{RA} is Space Complexity Recursive Algorithm for recursive calls and S_{RACUM} is Cumulative count for space complexity for recursive algorithm which is not applicable.

Also, A = $\{A_i\ A_j A_k\}$ where the three key components are $A_i, A_j\ and\ A_k$ called Algorithm line number, Algorithm statement and Frequency count respectively. Where A equals sets of key components identified. A_i = 1, 2…n, where A_j = Statement 1, Statement 2…, Statement which represents the computations performed at each line of algorithm. A_k = Count1, Count2 …, Countn which is computations performed at each Algorithm line

Note: Counti: is the number of times each operation is executed. A = $\{A_i, A_j, A_k\}$ Acum= $\{A_{SCCUM}, A_{SVCUM}, A_{SDCCUM}\}$

Therefore, the cumulative counts can be expressed as S_{CCUM} = $\{A_{SCCUM} = \sum_{i=1}^{K} Asci\}$, S_{VCUM} = $\{A_{VCUM} = \sum_{i=1}^{K} Asvi\}$, S_{DCCUM} = $\{A_{DCCUM} = \sum_{i=1}^{K} A_{DCCUMi}\}$. Where k, represents the number of operations associated with each algorithmic line.

This structured approach to the implementation of the EMHES ensures that both time and space complexities are thoroughly analyzed. Hence, an efficient and scalable system design and develop.

5. Result and Performance Analysis

5.1 The Complexity Analysis Came From the Depicted Flowchart

Figure 1 represents the Flow chat of the developed architecture where the process begins with the retrieval of the Vulnerability Score (VS), which is digitally signed to ensure integrity and authenticity. The system then determines the number of participating parties for the multi-party computation, after which cryptographic keys are generated and distributed among them. These keys are employed for encryption and decryption operations within the homomorphic encryption (HE) framework. A verification process is applied in which the system confirms whether the distributed key has been successfully generated and accepted by all parties. If agreed and key exchange are successful, the process continues otherwise, it iterates until all keys are correctly generated and distributed. Each party (P1 to Pn) privately retains its decryption key. The VS is then encrypted using HE to form Encrypted Vulnerability Score (EVS). Homomorphic operations are applied directly to the EVS, and the resulting ciphertext is transmitted to the Cloud Service Provider (CSP) where computations and storage are outsourced. During decryption stage, the digital signature of the ciphertext is first verified to confirm that the EVS has not been tampered during the transmission or processing operations. The system again determines the number of parties involved in the decryption phase and reconstructs the secret decryption key. If the required decryption keys are reconstructed and confirmed, the process proceeds; otherwise, it repeats until the decryption key can be successfully reconstructed. Lastly, the EVS is decrypted using the secret key, and the original VS is recovered with its digital signature intact, thereby maintaining both integrity and authenticity.

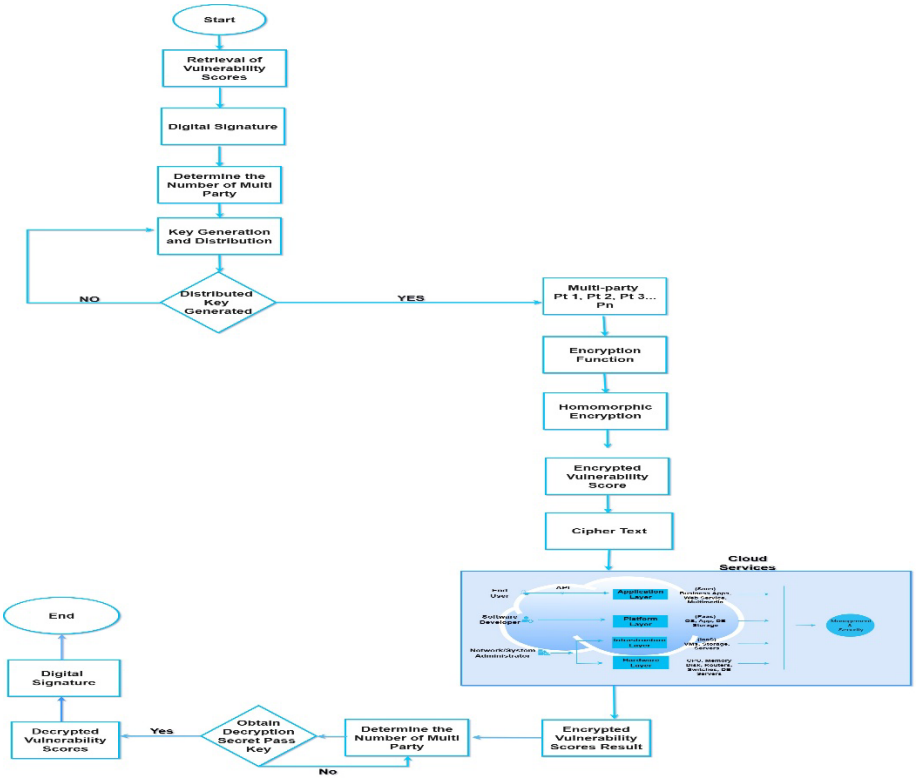


Figure 1: Flowchart for the Developed Architecture

5.2 Architecture for an Enhance Multiparty Homomorphic Encryption Scheme (EMHES)

Figure 2 represents EMHES. The architecture enables privacy-preserving computation and secure outsourcing of vulnerability scores (VS) to an untrusted cloud environment using homomorphic encryption. Vulnerability Scores (VS) are obtained from classified data, which are digitally signed to ensure integrity and authenticity. The digitally signed VS scores are handled by a trusted third-party government regulatory agency, where the data and keys are managed in a secure environment. Homomorphic encryption is initiated by the system where encryption keys are generated, and a multi-party computation (MPC) setup is established among participating entities (P1 to Pn). Each party receives and keep its own decryption key share, ensuring that no single party can decrypt the data unilaterally. The VS are encrypted using a homomorphic encryption scheme to produce Encrypted Vulnerability Scores (EVS), while preserving the ability to perform computations directly on the ciphertext. EVS is transmitted via a secured transmission channel to the Cloud Service Provider (CSP), where storage and computation are outsourced. Within the cloud environment, various layers such as application layer platform layer, Infrastructure layer etc. operates over the EVS without accessing to the plaintext VS data. These operation leverages on homomorphic function, allowing computations to be performed on ciphertext while data confidentiality is maintained. Note that the digital signatures attached to the EVS continue to protect against tampering, ensuring that any unauthorized alteration can be detected. During decryption stage, the EVS are returned via a secured transmission channel to the trusted environment where the digital signatures is verified and confirm that the EVS has not been altered during cloud processing or transmission. It then re-engages the multi-party decryption process, where the set of parties required for key reconstruction are determined and secret key are shares. The EVS is subsequently decrypted and decrypted VS retain their digital signatures, thereby preserving integrity and authenticity from end to end.

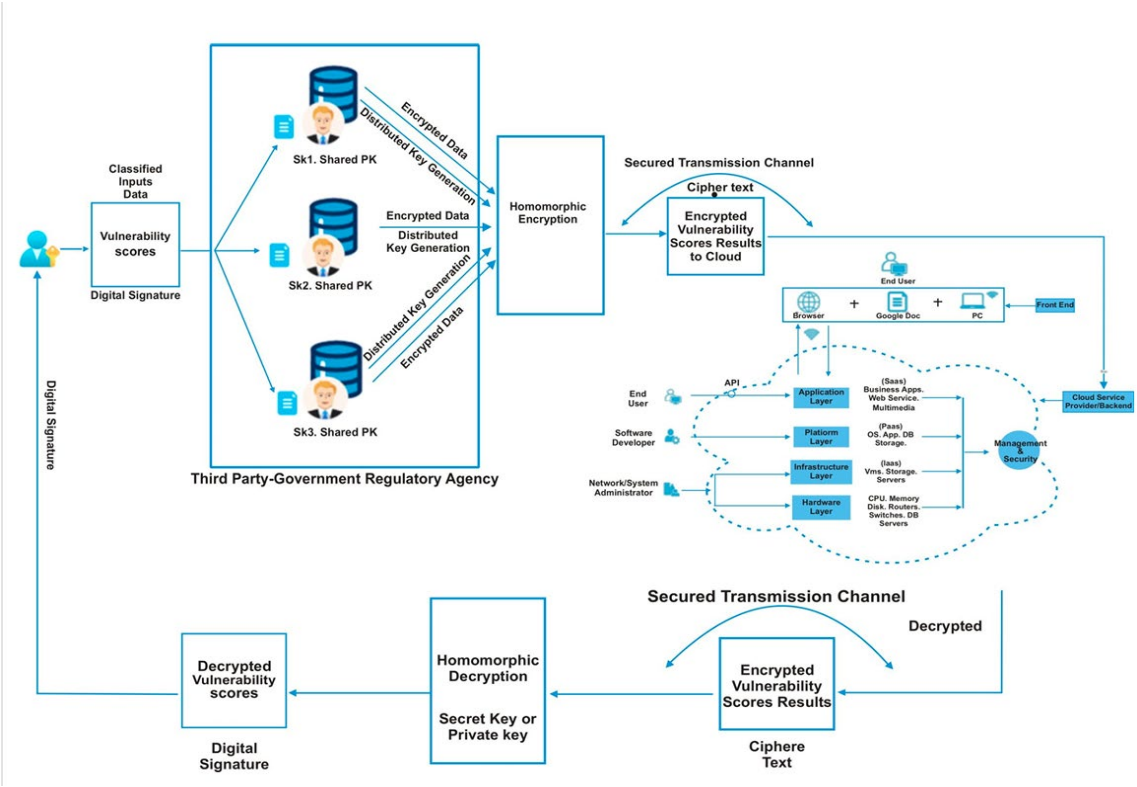


Figure 2: Architecture for an Enhance Multiparty Homomorphic Encryption Scheme (EMHES)

5.3 Result of the Analysis of Algorithm

Table 1 shows a comparative analysis of time complexities of algorithm for the designed architecture and baseline algorithms

Table 1: EMHES analysis

S/N	Author	Result of Time Complexity	Time complexity chat	Percentage Comparative Analysis
1	(Bao et al. 2024)-Base Line 1	O(1) + 32	Constant	0.32%
2	(Zhang et al. 2022)-Baseline 2	O(1) +101	Constant	0.98%
3	EMHES Architecture	6n+72	Linear complexity	6.52%

Table 1 shows the analysis of the algorithm od designed architecture (EMHES) for thetime complexity against existing algorithms. The EMHES algorithm exhibits a time complexity of 6n + 72 (O(n)), indicating that its computational effort grows linearly with the input size while the baseline algorithms from (Bao et al. 2024) and (Zhang et al. 2022) show constant time complexities of O(1) + 32 and O(1) + 101, respectively. Additionally, the EMHES architecture demonstrates a comparative analysis percentage of 6.52%, significantly higher than the 0.32% and 0.98% of the baseline algorithms. This indicates that the EMHES architecture offers enhanced security against Man-in-the-Middle (MitM) attacks, as it mandates increased computational effort from potential attackers, thereby improving overall security resilience.

Key assumptions:

- 1. General format of Constant Time Complexity of an algorithm (T_c):

$T_c = O(1) + K$

Where K is the units of time of algorithmic execution and O(1) is a Unit Big-O time of algorithmic execution

- 2. General format of Linear Time Complexity of an algorithm (T_L):

$T_L = O(n) + K$

Where O(n) is the n-time linear Big-O time of algorithmic execution and $K \leq n < \infty$

- 3. General format of Quadratic Time Complexity of an algorithm (T_Q):

$T_Q = O(n^2) + O(n) + K$

Where $O(n^2)$ is the n^2-time quadratic Big-O time of algorithmic execution

- 4. In this designed algorithmic analysis, let us assume n = K,

Where K is the greatest unit of time of algorithmic execution,

i.e., n = 101

Percentage (%) Comparative Analysis of the developed algorithm against the baseline work are given below.

For the Designed/Developed Algorithm (DA)

Note: Based on the assumption, n = 101

% PA Comparative Analysis

$(\%PA_{CA}) = \frac{Linear\ Time\ Complexity}{Quadratic\ Time\ Complexity}\ X\ 100$

$\%PA_{CA} = \frac{6n+72}{n^2+n+K}\ X\ 100$

$\%PA_{CA} = \frac{6(101)+72}{101^2+101+101}\ X\ 100$

$\%PA_{CA}$ = 6.52%

For the Baseline1 Algorithm (B1A),

% B1A Comparative Analysis

$(\%B1A_{CA}) = \frac{Constant}{Quadratic\ Time\ Complexity}\ X\ 100$

$\%\ B1A_{CA} = \frac{O(1)+32}{101^2+101+101}\ X\ 100$

$\%\ B1A_{CA} = \frac{1+32}{101^2+101+101}\ X\ 100$

$\%\ B1A_{CA} = \frac{33}{101^2+101+101}\ X\ 100$

$\%\ B1A_{CA}$ = 0.32%

For the Baseline2 Algorithm (B2A),

% B2A Comparative Analysis

$(\%B2A_{CA}) = \frac{Constant}{Quadratic\ Time\ Complexity}\ X\ 100$

$\%\ B2A_{CA} = \frac{O(1)+101}{101^2+101+101}\ X\ 100$

$\%\ B2A_{CA} = \frac{1+101}{101^2+101+101}\ X\ 100$

$\%\ B2A_{CA} = \frac{102}{101^2+101+101}\ X\ 100$

$\%\ B2A_{CA}$ = 0.98%

Hence $\%PA_{CA}$ = 6.52%, $\%\ B1A_{CA}$ = 0.32%, and $\%\ B2A_{CA}$ = 0.98%

Figure 3 shows the Time Complexity for The Designed Architecture. The bar chart which displays the relationship between Algo line number and number of frequency(ies) for 70 data points representing lines of code. It processes more input and shows that as the algorithm's line number increases. The proposed model (EMHES) incorporating Big O notation, demonstrates superior security, integrity, and performance compared to baseline models without Big O, and is better equipped to handle larger datasets.

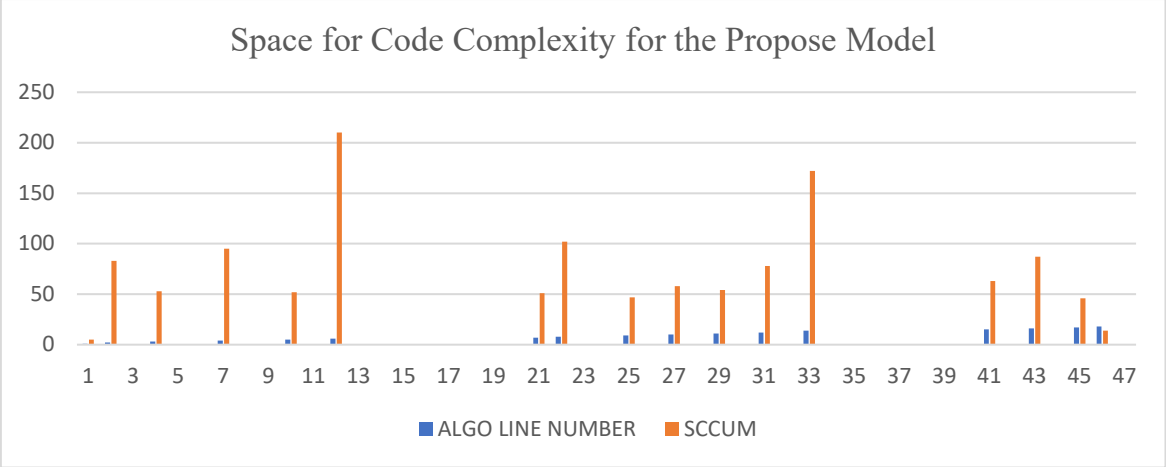


Figure 3: Time Complexity

C. Comparative analysis of space complexities of the algorithm

The comparative analysis of space complexities for the Enhanced Multi-Party Homomorphic Encryption Scheme (EMHES) against baseline algorithms is shown through a series of figures (Figures 4 to 5). These figures illustrate the space complexity of the designed architecture, showing metrics such as space for code complexity and data structure. Space Complexity Recursive Algorithm is not applicable in our algorithm as never recall itself with its on definition.

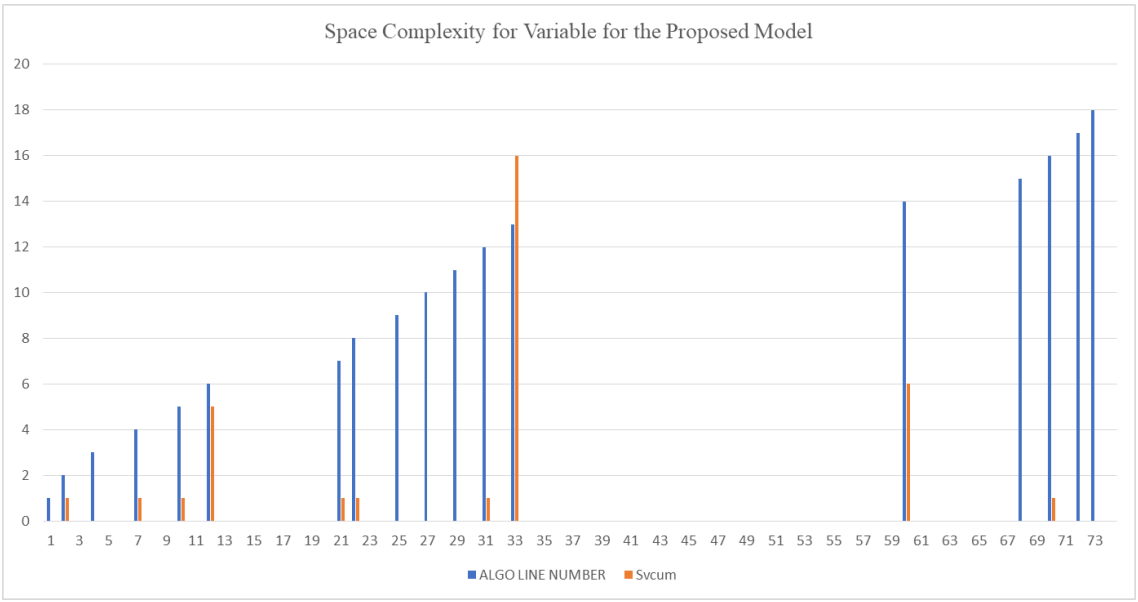


Figure 4. Space for Code Complexity

Figure 4 shows a Algo line number and Space for code complexity cumulative. Both plotted against an x-axis ranging from 1 to 73. The y-axis, representing space complexity cumulatively extends from 0 to 1400. S_{CCUM}, which represents the total memory space consumed by the algorithm during execution. This shows a significant spike around x-axis values 11, 33, and 60, at peak of 1300-1350 where x=33. These spikes indicate considerably higher memory resource usage, suggesting that certain algorithm computations, particularly due to cloud activities at x=33 and x=60, demand substantial computational resources and memory.

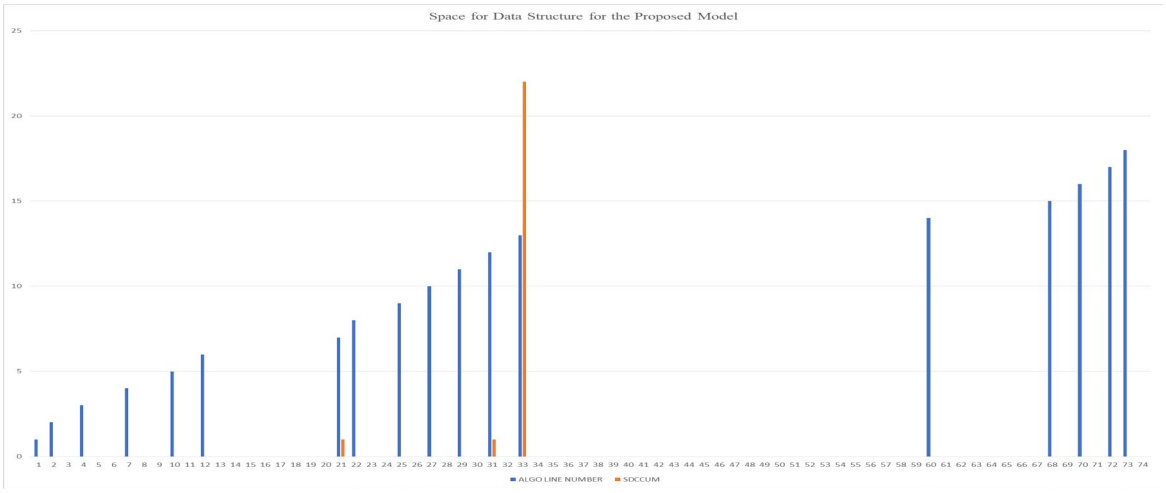


Figure 5: Space complexity for Data Structure EMHES

Fig. 5 illustrate the space complexity for data structures of the designed algorithm, with the x-axis representing Algorithm line number ranging from 1 to 73 and the y-axis showing space complexity ranging from 0 to 25. The Algo Line Number which indicate space complexity for data structures at various algorithm lines, while the cumulative space complexity data structure (S_{DCCUM}) values at each algorithm line. The Algo Line Number peaks around lines 59-73 where it has values of 14 and 18, whereas S_{DCCUM} has a significant peak algorithm line 33, reaching 22.

5.4 Discussion

This research provides comparisons of EMHES with existing frameworks of (Bao et al. 2024) "Direct Homomorphic Secured Multi-party computation (DHSMPC)", (Zhang et al. 2022) "Multi-party Quantum Homomorphic Encryption (MQHE)" and (Sanon et al. 2023) "Secure Federated Learning with HE". While DHSMPC focuses reduction of ciphertext size in for efficient storage and communication, it fails to address the complexity growth with larger input sizes. In contrast, EMHES demonstrates linear complexity, ensuring practical operation and enhancing resistance to MitM attacks as datasets increase. EMHES also offers immediate applicability in current cloud and regulatory infrastructures, where quantum solutions are not yet widely implemented. The work of (Sanon et al. 2023) highlights the trade-off between security and computational overhead in FL with homomorphic encryption, revealing that cryptographic protections can significantly extend processing times. EMHES mitigates this issue as it employs controlled linear time complexity, providing both scalability and security without excess overhead. Moreover, EMHES shows secure and tamper-resistant vulnerability evaluation, critical in regulatory settings where data integrity and authenticity must be maintained. EMHES provides a scalable solution that maintains performance and enhances attack resistance, making it ideal for large-scale, integrity-critical cloud-based vulnerability management, unlike existing studies that mainly emphasize compactness, quantum resilience, or privacy preservation.

6. Conclusion

In this study, EMHES was introduced as a robust architectural framework designed to mitigate MitM attacks on vulnerability score (VS) manipulation. Homomorphic encryption is integrated with advanced security measures namely digital signature and key management protocols, EMHES clearly preserve the confidentiality, integrity and authenticity of the VS through the process. In computational analysis, EMHES architecture maintains a linear time complexity contrasting significantly with the existing algorithm that shows constant time complexity. This attributes not only enhances security resilience against any potential attackers but also accommodate larger datasets, making it scalable. Additionally, the performance metrics based as assumed *n=100* demonstrated Comparative Analysis score of 6.25% for EMHES, evidently outperforming baseline algorithms. Furthermore, the detail analysis of the space complexity shows that the algorithm line of code requires significant computational resources especially during the cloud processing, showing areas of potential optimization. In conclusion, the EMHES framework offers a comprehensive, secure and efficient mechanism for collaborative processing of data is sectors where integrity and confidentially are important such as healthcare, finance and cybersecurity sectors.

Acknowledgements

All the technical inputs from team members and partners are all acknowledged and appreciated.

Ethics declaration: There is no ethical bridge or conflict in this research work necessitating ethical clearance and declaration. All actions complied with research guidelines as stipulated by the conference guidelines.

AI declaration: We employed POE to assist in conducting few literatures review. POE helped in identification of the relevant studies, but the selection and all the interpretation of the literature were performed by the authors.

References

Adablanu, S, Potter, K, Stilinki, D & Stilinski, D 2024, 'Homomorphic Encryption for Secure Cloud Computing Homomorphic Encryption for Secure Cloud Computing Homomorphic Encryption for Secure Cloud Computing'.

Alanazi, M, Mahmood, A & Chowdhury, MJM 2023, *SCADA vulnerabilities and attacks: A review of the state-of-the-art and open issues*, Elsevier Ltd, *Computers and Security*, vol. 125.

Appan, A, Chandramouli, A & Choudhury, A 2023, 'Network Agnostic Perfectly Secure MPC Against General Adversaries', in *Leibniz International Proceedings in Informatics, LIPIcs*, vol 281, Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing.

Bao, H, Yuan, M, Deng, H, Xu, J & Zhao, Y 2024, 'Secure multiparty computation protocol based on homomorphic encryption and its application in blockchain', Elsevier Ltd, *Heliyon*, vol. 10, no. 14.

Chen, Q, Li, H, Ariffin, S & Abdullah, NAS 2024, 'Overview of homomorphic encryption technology for data privacy', Conscientia Beam, *Review of Computer Engineering Research*, vol. 11, no. 3, pp. 130–139, viewed 6 September 2025, <https://ideas.repec.org/a/pkp/rocere/v11y2024i3p130-139id3955.html>.

Das, S, Ray Chowdhury, S, Chandran, N, Gupta, D, Lokam, S & Sharma, R 2025, 'Communication Efficient Secure and Private Multi-Party Deep Learning', Privacy Enhancing Technologies Symposium Advisory Board, *Proceedings on Privacy Enhancing Technologies*, vol. 2025, no. 1, pp. 169–183.

Ghanem, SM & Moursy, IA 2019, 'Secure Multiparty Computation via Homomorphic Encryption Library', Institute of Electrical and Electronics Engineers Inc., *Proceedings - 2019 IEEE 9th International Conference on Intelligent Computing and Information Systems, ICICIS 2019*, pp. 227–232.

Hashemi Chaleshtori, F & Ray, I 2023, 'Automation of Vulnerability Information Extraction Using Transformer-Based Language Models', Springer, Cham, *Lecture Notes in Computer Science*, vol. 13785 LNCS, pp. 645–665, viewed 7 September 2025, <https://link.springer.com/chapter/10.1007/978-3-031-25460-4_37>.

Jiang, Y, Oo, N, Meng, Q, Lim, HW & Sikdar, B 2025, 'A Survey on Vulnerability Prioritization: Taxonomy, Metrics, and Research Challenges', *Cornell University*, vol. 1, viewed 7 September 2025, <https://arxiv.org/pdf/2502.11070>.

Kalouptsoglou, I, Siavvas, M, Ampatzoglou, A, Kehagias, D & Chatzigeorgiou, A 2024, 'Vulnerability prediction using pre-Trained models: An empirical evaluation', IEEE Computer Society, *Proceedings - IEEE Computer Society's Annual International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems, MASCOTS*.

Kamble, A, Jiet, MM & Puri, C 2024, 'Homomorphic Encryption and its Applications in Multi-Cloud Security', Institute of Electrical and Electronics Engineers Inc., *7th International Conference on Inventive Computation Technologies, ICICT 2024*, pp. 1493–1499.

Keskin, O, Gannon, N, Lopez, B & Tatar, U 2021, *XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE Scoring Cyber Vulnerabilities based on Their Impact on Organizational Goals**.

Li, S, Li, S, Cai, X-Q, Cai, X-Q, Cai, X-Q, Wang, T-Y, Wang, T-Y & Wang, T-Y 2025, 'Secure multiparty computation for maximum and minimum values based on quantum homomorphic encryption', Optica Publishing Group, *Optics Express, Vol. 33, Issue 7, pp. 16263-16274*, vol. 33, no. 7, pp. 16263–16274, viewed 7 September 2025, <https://opg.optica.org/viewmedia.cfm?uri=oe-33-7-16263&seq=0&html=true>.

Mirtaheri, SL, Pugliese, A, Movahed, N & Shahbazian, R 2025, 'A comparative analysis on using GPT and BERT for automated vulnerability scoring', Elsevier B.V., *Intelligent Systems with Applications*, vol. 26.

Patil, S & Patra, A 2025, 'Perfectly-secure Network-agnostic MPC with Optimal Resiliency', Association for Computing Machinery (ACM), pp. 121–130, viewed 7 September 2025, </doi/pdf/10.1145/3732772.3733500?download=true>.

Pentyala, S, Railsback, D, Maia, R, Dowsley, R, Melanson, D, Nascimento, A & De Cock, M 2022, 'Training Differentially Private Models with Secure Multiparty Computation'.

da Ponte, FRP, Rodrigues, EB & Mattos, CLC 2023, 'CVEjoin: An Information Security Vulnerability and Threat Intelligence Dataset', Springer, Cham, *Lecture Notes in Networks and Systems*, vol. 661 LNNS, pp. 380–392, viewed 7 September 2025, <https://link.springer.com/chapter/10.1007/978-3-031-29056-5_34>.

Sanon, SP, Reddy, R, Lipps, C & Schotten, HD n.d., *Cross-Silo Horizontal Federated Learning Methods in Network Traffic Analysis*.

Sanon, SP, Reddy, R, Lipps, C & Schotten, HD n.d., *Secure Federated Learning: An Evaluation of Homomorphic Encrypted Network Traffic Prediction*.

Song, C, Huang, R & Hu, S 2024, 'Private-preserving language model inference based on secure multi-party computation', Elsevier B.V., *Neurocomputing*, vol. 592.

Tom & Kosacki 2023, 'BT_Prioritising_Vulnerability_Scan_Results'.

Yan, G, Zhang, Y, Guo, Z, Zhao, L, Chen, X, Wang, C, Wang, W, Meng, D & Hou, R 2025, 'Comet: Accelerating Private Inference for Large Language Model by Predicting Activation Sparsity', Institute of Electrical and Electronics Engineers Inc., *Proceedings - IEEE Symposium on Security and Privacy*, pp. 2827–2845.

Yuan, S, Shen, M, Mironov, I & Nascimento, ACA 2021, 'Practical, Label Private Deep Learning Training based on Secure Multiparty Computation and Differential Privacy', *Cryptology ePrint Archive*, viewed 7 September 2025, <https://eprint.iacr.org/2021/835>.

Zeng, W, Dong, Y, Zhou, J, Ma, J, Tan, J, Wang, R & Li, M 2025, 'MPCache: MPC-Friendly KV Cache Eviction for Efficient Private Large Language Model Inference', viewed 7 September 2025, <http://arxiv.org/abs/2501.06807>.

Zhang, JW, Chen, XB, Xu, G, Li, HJ, Wang, YL, Miao, LH & Yang, YX 2022a, 'A Secure Multiparty Quantum Homomorphic Encryption Scheme', Tech Science Press, *Computers, Materials and Continua*, vol. 73, no. 2, pp. 2835–2848.

Zhang, JW, Chen, XB, Xu, G, Li, HJ, Wang, YL, Miao, LH & Yang, YX 2022b, 'A Secure Multiparty Quantum Homomorphic Encryption Scheme', Tech Science Press, *Computers, Materials and Continua*, vol. 73, no. 2, pp. 2835–2848.

Zhang, J-W, 张静文, Chen, X-B, 陈秀波, Xu, G, 徐刚, Yang, Y-X & 杨义先 2021, 'Universal quantum circuit evaluation on encrypted data using probabilistic quantum homomorphic encryption scheme*', IOP Publishing, *Chinese Physics B*, vol. 30, no. 7, p. 070309, viewed 7 September 2025, <https://iopscience.iop.org/article/10.1088/1674-1056/ac003b>.

Zhou, J, Feng, Y, Wang, Z & Guo, D 2021, 'Using secure multi-party computation to protect privacy on a permissioned blockchain', MDPI AG, Sensors, vol. 21, no. 4, pp. 1–17.

Blockchain Use in Disbursing Financial Aid at Higher Education Institutions

Livhuwani Mathintha, Noluntu Mpekoa and Tevin Moodley
University of Johannesburg, South Africa

217034673@student.uj.ac.za
noluntum@uj.ac.za
tevin@uj.ac.za

Abstract: Financial aid plays a crucial role in ensuring access to education, particularly for individuals from disadvantaged financial backgrounds. Education is a powerful driver of cultural and national development. By facilitating education, financial aid cultivates young professionals capable of driving progress across various fields. Notably, approximately 84 percent of students benefit from financial aid scholarships. In South Africa, bursaries and scholarships play a crucial role in providing essential financial support to students. However, the existing disbursement process is fraught with significant challenges. The manual nature of these processes leads to time-consuming operations, a high risk of errors, delays in fund distribution, and vulnerability to fraud. These inefficiencies hinder the timely delivery of financial aid, posing severe risks to students' academic success. To address these pressing challenges, it is imperative to transition to automated systems that streamline operations, minimise errors and delays, and enhance security against fraud. This transformation will ensure that financial aid reaches deserving students efficiently and securely. Therefore, this study conducts a systematic literature review to thoroughly investigate the challenges faced by financial aid institutions and students in fund disbursement at higher education institutions. Additionally, it examines the potential of Blockchain technology in improving fund disbursement processes. Findings indicate that the current scholarship disbursement procedures in South African institutions are manual, inefficient, and prone to errors, resulting in delays and opportunities for fraud. The urgency of this issue cannot be overstated, as students suffer from both hunger and frustration, while some lose their funds altogether. Importantly, the research highlights that a Blockchain ledger is inherently challenging to tamper with, thereby providing enhanced security. In a Smart Contract Blockchain-based system, every transaction is meticulously recorded on a distributed, publicly visible ledger, ensuring traceability and transparency of all disbursements and significantly reducing the risk of fraud.

Keywords: Blockchain, Smart contracts, Financial aid disbursements, Transparency, Education

1. Introduction

Financial aid plays a crucial role in ensuring access to education, particularly for individuals from disadvantaged financial backgrounds (Hutchinson, 2025). “Education remains a fundamental drive of social, cultural, and national development” (Alshamaila, 2024). Bouchrika (2024) states that education helps in equipping individuals with skills and knowledge that fuel innovation, productivity, and economic growth. By facilitating education through financial support, financial aid programs help cultivate skilled professionals who can make meaningful contributions to various industries and sectors, thereby fostering innovation and development. As reported by Bouchrika (2024), the average published tuition and fees for full-time students in the academic year 2023-24 have increased by 2.5 percent to 4 percent from the previous year. The study also revealed that approximately 84 percent of students receive financial aid through need-based grants. Bursaries and scholarships are critical, as they not only provide support that empowers individuals but also foster inclusivity and equal opportunity within the education system. This provision also contributes to the reduction of inequality and the enhancement of long-term socio-economic outcomes (Bouchrika, 2024). However, the current manual financial aid disbursement process faces major challenges, including administrative delays, human errors, lack of transparency and vulnerability to fraud, which hinder the timely and fair distribution of funds (Tressa & Priya, 2024). These inefficiencies negatively affect students’ academic progress and trust in the system. To address this, a shift toward automated, secure, and transparent solutions is essential (Bedi et al, 2020). This study aims to investigate the challenges faced by financial aid institutions and students in fund disbursement at higher education institutions.

By examining these challenges in depth, the study highlights the barriers that limit the effectiveness of financial aid systems and their ability to meet the needs of students equitably and efficiently. The study also aims to examine the potential of Blockchain technology in improving fund disbursement processes. The study explores how Blockchain features, such as decentralisation, immutability, and automation through Smart Contracts, could address current inefficiencies and enhance transparency, security, and accountability in financial aid management. This study is essential as it contributes to both academic knowledge and methods for integrating Blockchain into financial aid processes. The study is timely given the increasing demand for transparency, accountability, and speed in financial transactions across educational sectors (Shahrukh, 2023). This analysis

could provide insights for global institutions, highlighting potential approaches and considerations for creating a more inclusive higher education.

This paper is organized as follows: Section 2 outlines the study background and identifies a research gap in the disbursement of financial aid at higher education institutions, highlighting the potential for Blockchain solutions. Section 3 details the research methodology, including the strategies employed to investigate challenges in financial aid disbursement and develop a suitable solution. Section 4 presents the results from the literature review and analysis. Section 5 discusses the findings, focusing on their significance, practical implications, and the trade-offs between centralized efficiency and decentralized trust, as well as technical and economic limitations like asset volatility and Smart Contract rigidity. Section 6 offers recommendations for implementing Blockchain solutions and suggests strategies to address scalability and regulatory challenges. Finally, Section 7 concludes the study by summarizing key insights and contributions.

2. Study Background

2.1 Traditional Fund Disbursement Systems

Bedi et al. (2020) identified key issues in India's Central Sector Scheme of Scholarship (CSS), such as poor traceability of applications, lost forms during postal transit, limited transparency, and bank account verification challenges. Similarly, Tekguc, Adalier, and Yurtkan (2020) highlighted problems in Turkey's traditional scholarship management at the Credit and Hostels Institution (CHI), including manual verifications, inefficiencies, and high maintenance costs, alongside potential vulnerabilities in Smart Contracts. Bálint et al. (2019) noted issues in traditional bursary systems, such as a lack of transparency and reliance on intermediaries. Jadhav and Pise (2023) also found problems like poor communication and tracking difficulties in scholarship management. Honesti et al. (2022) discussed time-consuming manual processes and delays in the Gamification Scholarship Scheme. Despite multiple studies addressing specific challenges, no research has compiled a comprehensive list of issues faced by financial aid institutions and students in fund disbursement in higher education. This study aims to fill that gap.

Delays, fraud, and lack of transparency in financial aid significantly impact student performance and well-being. At Kenyatta University, 17% of students faced HELB loan disbursement delays, leading to disrupted attendance and poorer academic outcomes; students without delays had GPAs 0.647 points higher (Borner et al., 2022). In South Africa, cyber fraud has caused severe financial losses for 100% of affected financial aid recipients, with many losing over R5,000 and none recovering their funds. This resulted in academic disruptions, with 27% experiencing grade deterioration and 35% dropping out, alongside psychological distress for 93% and loss of trust in institutions for 90% (Hutchinson, 2025). Additionally, a lack of transparency leaves 61% of first-generation and 40% of low-income students confused and at risk of increased debt, leading to lower enrollment and higher dropout rates. Blockchain-based systems are proposed as a solution to enhance transparency and accountability in financial aid disbursement (Rane, Choudhary, & Rane, 2023).

2.2 Technology-advanced Fund Disbursement Systems

Bedi et al. (2020) implemented a Smart Contract-based system on Ethereum for automating scholarship authentication, enhancing transparency, and preventing application loss, though it remained vulnerable to Denial-of-Service and Man-in-the-Middle attacks, while overlooking data privacy and regulatory compliance. Tekguc, Adalier & Yurtkan (2020) introduced the ScholarChain platform in Turkey, which used Blockchain and Smart Contracts for document verification and scholarship processes. Despite its integration of Oracle services for data accuracy, challenges include reliance on external sources, high costs, and technical complexity. Jadhav & Pise (2023) presented CryptoScholarChain, a Blockchain framework improving scholarship management with Ethereum Smart Contracts and a ReactJS front end yet faced issues with Smart Contract complexity and scalability. Honesti et al. (2022) proposed a Blockchain-based solution for a Gamification Scholarship Scheme, automating application processes and fund disbursement; however, it still confronts accessibility challenges, including the need for mobile applications and system scalability.

The studies discussed above indicate that while Blockchain technology offers certain advantages, it also presents notable disadvantages. To date, no research has comprehensively explored and delineated the benefits and potential challenges of implementing Blockchain in fund disbursement processes. Therefore, this study seeks to investigate the challenges faced by financial aid institutions and students in the distribution of funds at higher education institutions, and to examine how Blockchain technology could improve these processes.

2.3 Contribution of This Study

Based on this investigation, the authors' contribution lies in demonstrating the potential of Blockchain technology to improve the financial aid disbursement process. The analysis also assesses whether it constitutes an appropriate and effective solution to the challenges inherent in existing financial aid systems. This study examines existing scholarship management solutions and frameworks that leverage Smart Contracts to automate verification and fund transfers, thereby reducing delays and human error. The study also highlights practical considerations for implementing Blockchain in higher education, including system security, scalability, and integration with existing institutional processes.

3. Research Methodology

This paper seeks to investigate the challenges faced by financial aid institutions and students in fund disbursement at higher education institutions. The study also aims to examine the potential of Blockchain technology in improving fund disbursement processes. Therefore, this study conducts a systematic literature review to thoroughly investigate these challenges faced by financial aid institutions and students in fund disbursement. Systematic literature reviews (SRs) are a way of synthesising scientific evidence to answer a particular research question in a way that is transparent and reproducible, while seeking to include all published evidence on the topic and appraising the quality of this evidence (Lame, 2019). The use of a SR in this study will assist in facilitating a rigorous investigation of the challenges associated with fund disbursement in a transparent and unbiased manner. There are different systematic literature reviews; this study has utilised the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement. PRISMA was first established in 2009 and later updated in 2020 to ensure transparent and accurate reporting of systematic reviews (Matthew et al, 2021).

Table 1: Systematic Literature Review Approach

Database	Timespan	Keywords	First Round Results	Titles, abstracts, and conclusions	Research Aim
Google Scholar	2019-2025	Blockchain, Smart Contracts, financial aid disbursements, and education.	2056	125	23
EEE Xplore			19	5	3
ScienceDirect			23	8	4
			988	44	15

The main keywords used during the search were restricted to “Blockchain, Smart Contracts, financial aid disbursements, and education”. To capture as many relevant studies as possible, the study conducted electronic searches using Google Scholar, IEEE Xplore, and ScienceDirect databases, dating back to 2019 (Table 1). Articles and studies most relevant to the research theme were considered for this study. The searches on Google Scholar, ScienceDirect and IEEE Xplore in the first round yielded 2056 and 19 studies, 1011 respectively. For the ScienceDirect database, the author conducted two separate keyword searches. The first search, using the primary keyword, returned 23 records. A second search was conducted using the keywords Blockchain and Smart Contracts to broadly review the functionality of Blockchain technology yielding 988 articles. Articles not focusing on Blockchain-based financial aid disbursements, not written in English, published before 2019, or not peer-reviewed were excluded.

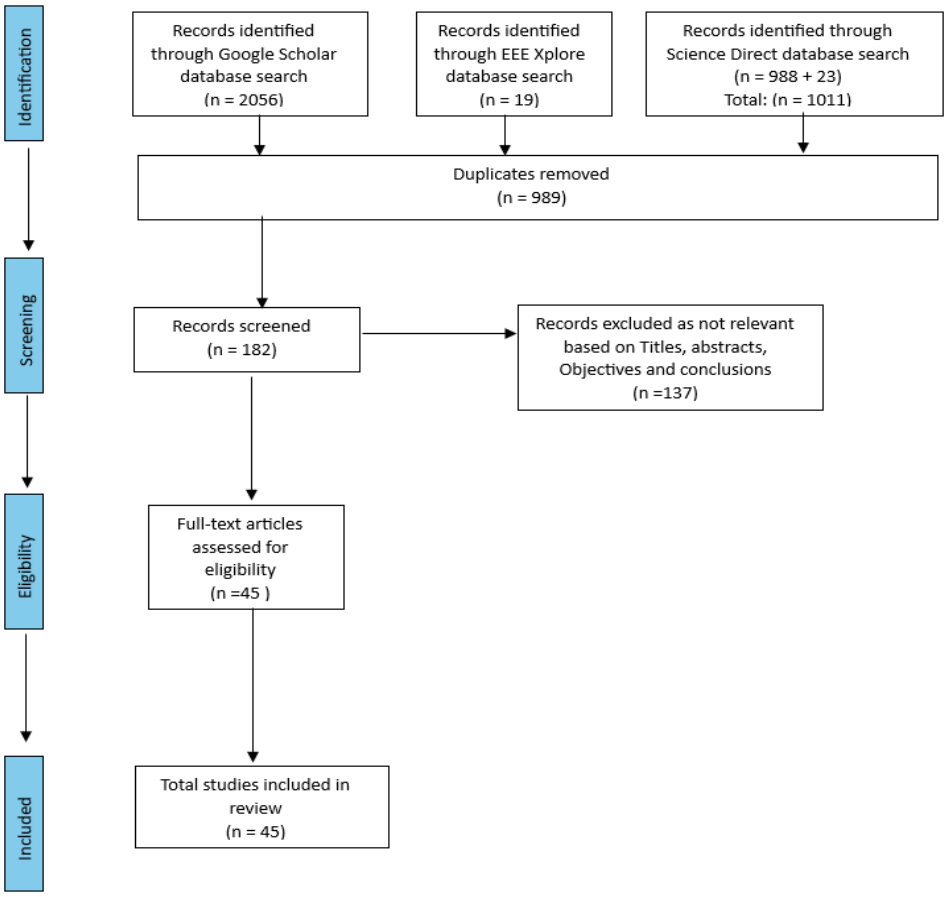


Figure 1: Literature search and filtering process

A total of 3086 journal articles and conference papers published between 2019 and 2025 were retrieved for this study. 989 duplicate records were removed and an initial screening of titles, abstracts, and conclusions resulted in 182 studies being shortlisted for further review. A subsequent full-text assessment was conducted to determine eligibility, and the studies that did not correspond with the research objectives were excluded. The primary objectives of this study are to investigate the challenges faced by financial aid institutions and students in the disbursement of funds at higher education institutions, and to examine the potential of Blockchain technology to enhance these processes. Following this evaluation, 45 studies were retained for qualitative analysis, as presented in Table 1 and Figure 1. The subsequent section provides a comprehensive discussion of the findings derived from this analysis.

4. Findings

The study by Davis (2022) highlights persistent shortcomings in the existing financial aid system, including overwhelming demand, administrative inefficiencies, and recurring delays in funds disbursement. To gain a deeper understanding of these issues, this review examined 40 scholarly articles. The analysis revealed recurring concerns regarding the transparency, efficiency, and timeliness of financial aid distribution, while also emphasising the potential of Blockchain technology as a viable solution to mitigate these challenges.

4.1 Challenges Faced in the Current Fund Disbursement

Bedi et al (2020) reported challenges faced by the Indian government in scholarship disbursement, including a lack of transparency, limited traceability, insecure verification, form loss, manual processing, and difficulties with bank account validation. Similar studies in financial aid and education have highlighted inefficiencies such as security risks and a lack of transparency in fund disbursement (Bedi et al, 2020; Honesti et al, 2022; Claudia, 2023), as presented in Table 2 below.

Table 2: Challenges in current financial aid disbursement

Authors	Challenges	
	Faced by universities	**Faced by students**
Bedi et al (2020)	Manual authentication, delays.	Lost forms, no transparency, no bank check, manual process.
Bálint et al (2019)	Error-prone systems, admin burden, weak security.	Manual paperwork causing delays.
Honesti et al (2022)	Manual checks, time-consuming.	Poor tracking and transparency.
Muneeb et al (2021)	Low security, Paper contracts.	Lack of trust.
Jadhav & Pise (2023)	Tamper-prone.	Tracking issues, lost documents, poor communication, and delayed process.
Asamoah et al (2023)	A long administrative process leading to a delay.	Long processing times or bureaucracy before funds are released.
Samu et al (2024)	Manual system processes leading to errors	Unfair and complex processes that lead to delays in releasing funds.
Tressa & Priya (2024)	Fraud, manual tasks, scalability.	Debt burden, no financial transparency.
Kibet (2025)	Data privacy risks, reliance on self-reports, poor cross-verification and delays.	Errors in self-reported data, fraud risk, and weak data security.
Solomon Christopher (2023)	Corruption, fraud, weak accountability.	Aid not reaching recipients, corruption, and outdated systems.
Emmanuel Goodluck (2025)	Loss of funds before they reach the intended beneficiaries	No transparency in fund use, fraud.
Jielong Yan et al (2025)	Privacy breaches, unreliable data, poor transparency.	Risk of unauthorised access to student data.
Nguyen-Hoang et al (2025)	Staff mismanagement of resources, excessive evaluation time and administrative resources	Privacy risks, data breaches, misconceptions, and false information.
Claudia (2023)	Inefficiency, fraud/corruption, weak accountability.	No transparency in fund use, low trust.

The study by Bedi et al (2020) also confirmed that students and financial aid institutions face multiple challenges in fund disbursement at higher education institutions. The study highlighted difficulties such as a lack of traceability of application forms, loss of forms during postal transit, limited transparency between students and education boards, and inadequate verification of bank accounts, which can lead to scholarship funds being returned if accounts are inactive. According to Jadhav & Pise (2023), traditional scholarship management systems are often centralised, lack transparency, rely on inefficient communication, are vulnerable to tampering, and face challenges in tracking and verifying applications. Similarly, Samu et al (2024) reported that the National Scholarship Portal (NSP) in India experiences challenges such as manual and complex application processes, physical verification requirements, delays in disbursement, excessive paperwork, and limited support services, reinforcing the notion that existing systems struggle with efficiency and reliability. Blockchain and Smart Contract technologies have been explored as potential solutions, with emphasis on enhancing trust, automating processes, and strengthening the financial aid disbursement system's resilience.

Blockchain technology has been identified as a potential solution to selected challenges in student financial aid systems, particularly fraud, lack of transparency, manual verification, and processing delays. Studies by Rane, Choudhary, and Rane (2023) indicate that the decentralised, cryptographic nature of Blockchain can reduce the risk of fraud by ensuring that transactions are securely recorded in an immutable, transparent ledger. Omar et al. (2020) highlighted that Blockchain enhances transparency and data integrity by improving data management and protocol compliance across participating entities. The use of smart contracts can automate eligibility

verification and fund disbursement processes, thereby reducing reliance on manual verification and minimising administrative delays. However, while Blockchain demonstrates strong potential in addressing security and trust-related issues, many of the other identified challenges can be effectively resolved through alternative centralised digital systems that offer automation, transparency, and reduced processing times without the complexity and overhead associated with Blockchain implementation.

4.2 Potential of Blockchain Technology

According to Sarmah (2018), Blockchain is a distributed ledger technology that records validated transactions across a global network of computers. Taherdoost (2023) defined Smart Contracts as automated programs within the Blockchain framework that enforce agreed-upon terms between parties. This review analysed scholarly works that apply Blockchain within educational and funding contexts, revealing both their strengths and vulnerabilities, and offering valuable insights for the development of a secure and efficient financial aid framework. Most studies incorporate Smart Contracts as a core component, and the review focuses on three key aspects (Table 3 below): the strengths, weaknesses, and applications of Blockchain in financial aid distribution.

Table 3: Strengths and Weaknesses of using Blockchain

Blockchain studies	Strengths in financial aid disbursement	Weaknesses in financial aid disbursement	Application / Focus
Bedi et al (2020)	Automation, transparency, verification, and error reduction.	Vulnerability to some cyberattacks, complexity, and limited testing.	Smart contract for scholarship disbursements
Bálint et al (2019)	Payment automation, transparency, security, low cost.	Security risks from connected systems.	Smart Contract for university bursary payments.
Honesti et al (2022)	Minimises human error, transparent transactions.	Cybersecurity vulnerabilities, legal and regulatory gaps.	Smart Contract-based gamification scheme for student selection, verification, multi-entity confirmation, and fund disbursement.
Muneeb et al (2021)	Robust security reduces costs and provides transparent implementation details.	Depending on external data sources.	Business transactions, many-to-many Smart Contracts.
Swati, J. & Nitin, P (2023)	Traceable transactions and efficient fund transfers.	High implementation costs, technical complexity, potential security risks from external components, and scalability limitations.	Ethereum-based Smart Contracts for the scholarship process in higher education, connecting donors, students, and NGOs.
Asamoah et al (2023)	Encryption, trust, privacy, and data integrity.	High transaction costs, latency, reliance on guarantors.	Blockchain-based student loans Smart Contracts.
Samu et al (2024)	Improve security, transparency, efficiency, inclusion, and simplify the scholarship application process.	Concerns over off-chain data security, lack of performance benchmarking.	Smart Contract to automate the scholarship portal to manage the entire scholarship lifecycle.
Zheng X.R. & Lu Y (2022)	Security enhancement, transparency, efficiency, and credential verification.	Cost, legal, adoption, and security challenges.	Secure financial operations, academic credentialing, and administrative automation.

Blockchain studies	Strengths in financial aid disbursement	Weaknesses in financial aid disbursement	Application / Focus
(Kibet 2025)	Improves transparency, security, and efficiency.	Relies on self-reported data and infrastructure challenges.	Need-based aid allocation.
Friday et al (2023)	Decentralised, tamper-proof records and real-time auditing.	Limited infrastructure, unclear regulations, scalability issues, and stakeholder resistance.	Automates fund release for public and international financial aid.
Goodluck (2025)	Lack of transparency, auditability, and accountability.	Resistance to change, policy and privacy uncertainties.	University funding.
Jielong Yan et al (2025)	Strong encryption, hierarchical access control, secure.	Prone to breaches and lack of adaptability.	Financial management Smart Contract-based system for educational institutions.
Tu-Anh Nguyen-Hoang et al (2025)	Lack of transparency preserves student privacy.	Low adoption.	Scholarship management platform.
Ngoc Cu Hoang et al (2024)	Ensures data integrity and immutability, reduces administrative overhead.	Scalability limits and latency.	Scalable Scholarship Management.

The analysis of the findings reveals noteworthy insights. Blockchain applications in financial aid consistently demonstrate strengths such as automation, transparency, decentralisation, enhanced security, trust, efficiency, and error reduction. These technologies have been applied to scholarship disbursements, bursary payments, student loan management, credential verification, university funding, and broader scholarship lifecycle management. However, recurring weaknesses include high implementation and transaction costs, scalability limitations, technical complexity, dependence on external data sources, cybersecurity vulnerabilities, and regulatory or policy gaps. The overall analysis highlights Blockchain's potential to transform financial aid systems by ensuring secure, transparent, and efficient fund management, while also emphasising the need to address challenges related to adoption, governance, and scalability for sustainable implementation.

5. Discussion

This section discusses the findings from the analysis, considering the core research objectives. This study was envisaged to investigate the challenges faced by financial aid institutions and students in the disbursement of funds at higher education institutions. This study also examined the potential of Blockchain technology in improving fund disbursement processes. The discussion also highlights the strengths and weaknesses of both traditional and Blockchain-based systems and interprets the lessons learned.

5.1 Challenges of Traditional Financial Aid Systems

The findings demonstrate that traditional financial aid systems face significant systemic and operational barriers that affect both institutions and students. Bedi et al (2020) identified critical inefficiencies in India's Central Sector Scheme of Scholarship (CSS), including manual verification of forms, delays in processing, and loss of documents during postal transmission. Honesti et al (2022) further highlighted that centralised platforms such as the National Scholarship Portal (NSP) often lack inclusivity, restricting access primarily to government-based scholarships and leaving private schemes unintegrated. Similarly, Kibet (2025) emphasised data integrity risks and privacy concerns in Kenya's New Funding Model (NFM), where opaque categorisation of students undermines fairness and creates mistrust. These studies collectively indicate that centralisation and reliance on manual processes are the root causes of inefficiency, opacity, and vulnerability to error or fraud. Administrative bottlenecks, inadequate bank account verification, and fragmented communication between agencies hinder the timely distribution of funds, undermining the primary purpose of student support schemes. Furthermore, the analysis by Samu et al (2024) reveals that existing scholarship systems, while attempting to centralise access, fail to deliver efficiency or equity due to structural issues. Authors, including Bedi et al (2020) and Kibet (2025),

highlight that the weaknesses stem from over-reliance on manual authentication, a lack of secure data sharing, and fragmented oversight. The analysis revealed that any future solution must directly address these inefficiencies by introducing automation, transparency, and secure stakeholder collaboration.

5.2 Potential of Blockchain Technology in Fund Disbursement

The analysis demonstrates that Blockchain technology, particularly through Smart Contracts, presents a viable and innovative solution to the challenges of traditional systems. Bedi et al (2020) proposed a Blockchain model that automates verification against educational databases, eliminating manual processes and reducing risks of form loss. Honesti et al (2022) emphasised that a Blockchain-based system could integrate private scholarships into a single ecosystem, enhancing inclusivity. Kibet (2025) demonstrated that Blockchain-based Smart Contracts can enhance data integrity and fairness in means testing by ensuring secure data exchange among multiple agencies. Similarly, Samu et al (2024) advocated for the adoption of permissioned frameworks such as Hyperledger Fabric, which protect privacy while maintaining trust across stakeholders. Analysis of prior studies indicates that Blockchain addresses the shortcomings of traditional systems by ensuring transparency through immutable records, fostering accountability through real-time fund tracking, and enhancing security against tampering or fraud (Claudia, 2023). Smart Contracts are a crucial mechanism here, as they enable automated disbursement once eligibility conditions are met, thereby reducing human error and delays. By removing the need for intermediaries, Blockchain-based Smart Contracts streamline processes, reduce administrative overhead, accelerate timelines, and lower transaction costs (Claudia, 2023). The capacity to integrate multiple scholarship sources within a decentralised system further supports fairness and accessibility.

Strengths: Blockchain introduces transparency, security, and efficiency into the fund disbursement process. Its tamper-proof nature enhances trust, while automation reduces administrative burden and accelerates fund release (Bedi et al, 2020; Samu et al, 2024). Private frameworks such as Hyperledger Fabric balance decentralisation with controlled access, making them suitable for sensitive financial and academic data (Samu et al, 2024). It was observed that Blockchain provides transparency, security, and efficiency in fund disbursement (Bedi et al, 2020). Smart Contracts were utilised by Bedi et al (2020) to automate processes, reducing administrative burdens and delays while eliminating intermediary costs. According to Claudia (2020), Kibet (2025) and Tressa & Priya (2024), decentralisation and immutability ensure tamper-proof records, improving trust and accountability. Additionally, cryptographic techniques safeguard sensitive student data, and the distributed ledger architecture enhances system resilience by eliminating single points of failure (Kibet, 2025). Permissioned frameworks, such as Hyperledger Fabric, preserve privacy and integrity (Challissery et al, 2024). Studies such as Tressa & Priya (2024) and Jadhav & Pise (2023) indicate that Blockchain enhances real-time transparency in tuition payments and funds. The study by Hoang et al (2025) utilised zk-rollups, which demonstrated significant scalability improvements by reducing transaction costs and confirmation times.

Blockchain technologies offer solutions for challenges in scholarship and financial aid systems, such as fraud, delays, and lack of transparency. Features like decentralized ledgers, smart contracts, and cryptographic methods can enhance transparency, automate fund disbursement, and improve data integrity. However, these advantages are limited by issues like dependence on accurate off-chain data, complex integration, privacy concerns, and ongoing security risks.

Weaknesses: Several limitations persist. Honesti et al. (2022) highlight the difficulty of accurately translating real-world processes into smart contract logic. Similarly, Claudia (2023) and Kibet (2025) identify high implementation costs, regulatory uncertainty, and the immutability of smart contracts as significant barriers, particularly when policy conditions change. Capacity building among stakeholders and the development of comprehensive legal frameworks for Blockchain technologies remain underdeveloped.

Beyond these general concerns, existing commercial Blockchain platforms that rely on smart contracts exhibit critical weaknesses, particularly in value-sensitive applications. A major issue is the extreme price volatility of crypto assets, which lack fundamental value and are largely driven by speculative market behaviour (Baur, Hong, & Lee). In student financial aid or scholarship disbursement, this volatility may result in beneficiaries receiving significantly more or less value than intended, thereby undermining budgeting accuracy, equity, and policy objectives (Yermack, 2024). Such instability raises serious concerns regarding the reliability, predictability, and accountability of Blockchain-based financial systems. The other factor is the technical complexity of smart contract development, which marginalises non-technical stakeholders, weakening claims of transparency and inclusivity. Smart contracts are susceptible to subtle programming vulnerabilities arising from discrepancies between developer intent and distributed execution semantics (Atzei et al, 2017). These risks are exacerbated by Blockchain immutability, as deployed contracts cannot be easily amended, often resulting in irreversible

financial losses (Atzei et al, 2017). Taken together, these economic and technical constraints underscore the limitations of Blockchain in socially and policy-critical applications.

Overall, this analysis indicates that the Blockchain solution must be adopted with caution. Its decentralised, secure, and automated characteristics may target the outlined challenges. Empirical prototypes and conceptual models proposed by Bedi et al. (2020) and Samu et al. (2024) further demonstrate the practical feasibility of Blockchain-based solutions. However, successful implementation depends on overcoming regulatory hurdles, managing technical complexity, and ensuring adequate stakeholder training. As several authors argue, Blockchain adoption represents not merely a technological upgrade but a systemic transformation requiring coordinated policy, institutional, and technological alignment.

Comparative analysis of Decentralised and Centralised solutions

This section evaluates decentralised Blockchain solutions compared to traditional centralised systems, addressing trade-offs in complexity, security, and practicality. The study by Islam et al. (2024) identifies strengths and weaknesses of each approach. Decentralised systems excel in security and transparency, as their distributed nature minimizes data tampering and enhances trust. However, they face scalability challenges. In contrast, centralised systems are efficient and scalable due to unified control but are prone to security vulnerabilities and reduced transparency, impacting trust (Islam et al. 2024). Blockchain can address security issues in multi-stakeholder environments like financial aid disbursement, where transparency is crucial. While centralised systems may streamline processes, Blockchain offers greater security and accountability despite increased complexity (Islam et al. 2024).

6. Recommendations and Future Work

This study reveals inefficiencies in traditional financial aid systems, including delays and a lack of transparency. It suggests that Blockchain technology can enhance these systems through decentralisation, immutability, automated Smart Contracts, and improved security. By incorporating Blockchain, scholarship and aid disbursements can become faster, more transparent, and secure, while also unifying government and private funding sources for greater inclusivity. Future research should address challenges such as scalability, regulatory issues, and the rigidity of Smart Contracts. Exploring stablecoin integration, hybrid models with legacy systems, and implementing machine learning-based fraud detection will be essential for a robust financial aid infrastructure. Leveraging these technologies can create a more equitable and efficient system.

7. Conclusion

This study investigated the challenges of financial aid disbursement and the potential of Blockchain to mitigate them. Findings reveal that traditional systems suffer from security vulnerabilities and administrative bottlenecks that hinder student success. Blockchain provides a viable alternative by enhancing transparency, accountability, and data integrity. However, practical implementation requires addressing system security, scalability, and the necessity of stakeholder training to bridge the technical gap. By integrating Blockchain with complementary technologies like machine learning, institutions can modernise financial aid into a secure, equitable system that ensures funds reach students efficiently.

Ethics declaration: This research did not require ethical clearance, as it did not involve human participants, animals, or sensitive personal data.

AI declaration: An AI tool (Grammarly) was utilised to improve grammar, spelling, and language clarity. The tool was not used for idea generation and analysis

References

Alshamaila, Y., Alsawalqah, H., Habib, M., Al-Madi, N., Faris, H., Alshraideh, M., Aljarah, I. and Masadeh, R. (2024) 'An intelligent rule-oriented framework for extracting key factors for grants scholarships in higher education', International Journal of Data and Network Science, 8(2), pp. 1325–1340.

Asamoah, K.O., Darko, A.P., Antwi, C.O., Kodjiku, S.L., Aggrey, E.S.E., Wang, Q. and Zhu, J. (2023) 'A Blockchain-based crowdsourcing loan platform for funding higher education in developing countries', IEEE Access, 11, pp. 24162–24174.

Atzei, N., Bartoletti, M. and Cimoli, T. (2017) 'A survey of attacks on Ethereum smart contracts', International Conference on Principles of Security and Trust, pp. 164–186.

Bálint, K., Cvetković, D., Takács, M., Holik, I. and Tóth, A. (2019) 'Connecting bitcoin Blockchain with digital learning chain structure in education', Acta Polytechnica Hungarica, 16(1), pp. 77–96.

Baur, D.G., Hong, K. and Lee, A.D. (2018) 'Bitcoin: Medium of exchange or speculative assets?', Journal of International Financial Markets, Institutions and Money, 54, pp. 177–189.

Bomer, A., Liu, X., Irungu, W.R. and Wanjiru, W.A. (2021) 'How is academic performance affected by delay in student loan disbursement in Kenyan universities? A case study of Kenyatta University', Higher Education Studies, 11(1), pp. 121–136.

Bouchrika, I. (2024) 'Financial aid and scholarships for online students in 2024', Education.

Claudia, M. (2023) 'Transforming Education Fund Management: Leveraging Blockchain for Transparency, Fraud Reduction, and Efficiency', Indonesia Auditing Research Journal, 12(4), pp. 179–186.

Friday, S.C., Lawal, C.I., Ayodeji, D.C. and Sobowale, A. (2023) 'Systematic review of Blockchain applications in public financial management and international aid accountability', International Journal of Multidisciplinary Research and Growth Evaluation, 4(1), pp. 1165–1180.

Goodluck, E. (n.d.) 'Higher education funding transparency through Blockchain: A secure data governance model for public universities'.

Hang'andu, N.B. (2022) 'Modes and challenges of accountability and transparency in the financial management systems in grant aided schools: A case of Kasisi girls and Kafue boys secondary schools' (Doctoral dissertation, The University of Zambia).

Hasan Shahrukh, M.R., Rahman, M.T. and Mansoor, N. (2023) 'A new paradigm in Blockchain-based financial aid distribution', in World Conference on Information Systems for Business Management, Cham: Springer, pp. 137–146.

Honesti, L., Aini, Q., Setiawan, M.I., Santoso, N.P.L. and Prihastiwi, W.Y. (2022) 'Smart contract-based gamification scheme for college in higher education', APTISI Transactions on Management (ATM), 6(2), pp. 102–111.

Hutchinson, G. (2025) 'The effects of cyber fraud on higher education financial aid students in South Africa'.

Islam, S. and Apu, K.U. (2024) 'Decentralized vs. centralized database solutions in Blockchain: Advantages, challenges, and use cases', Global Mainstream Journal of Innovation, Engineering & Emerging Technology, 3(4), pp. 58–68.

Khan, Y., Goyal, S.B. and Bedi, P. (2020) 'Security challenges of Blockchain', in International Conference on Innovations in Bio-Inspired Computing and Applications, Cham: Springer, pp. 248–257.

Kibet, A. (2025) 'Enhancing transparency and security in means testing for education financing: A Blockchain-based approach', in 2025 IST-Africa Conference (IST-Africa), IEEE, pp. 1–14.

Kshetri, N. (2017) 'Blockchain's roles in strengthening cybersecurity and protecting privacy', Telecommunications Policy, 41(10), pp. 1027–1038.

Luu, L., Chu, D.H., Olickel, H., Saxena, P. and Hobor, A. (2016) 'Making smart contracts smarter', Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 254–269.

Muneeb, M., Raza, Z., Haq, I.U. and Shafiq, O. (2021) 'Smartcon: A Blockchain-based framework for smart contracts and transaction management', IEEE Access, 10, pp. 23687–23699.

Nguyen-Hoang, T.A., Hoang, N.C., Hua, P.T., Thi, M.T.N., Ta, T.T., Nguyen, T., Tan-Vo, K., Dinh, N.T. and Nguyen, H.T. (2024) 'Advancing scholarship management: A Blockchain-enhanced platform with privacy-secure identities and AI-driven recommendations', IEEE Access.

Omar, I.A., Jayaraman, R., Salah, K., Simsekler, M.C.E., Yaqoob, I. and Ellahham, S. (2020) 'Ensuring protocol compliance and data transparency in clinical trials using Blockchain smart contracts', BMC Medical Research Methodology, 20(1), 224.

Rane, N., Choudhary, S. and Rane, J. (2023) 'Blockchain and artificial intelligence (AI) integration for revolutionising security and transparency in finance', SSRN 4644253.

Samu, L.C., Khithani, N., Alve, K., Gambhire, V., Hande, A. and Choubey, S. (2024) 'Blockchain based intelligent disbursement in National Scholarship Portal', IET Blockchain, 4(4), pp. 407–422.

Sarmah, S.S. (2018) 'Understanding Blockchain technology', Computer Science and Engineering, 8(2), pp. 23–29.

Swati, J. and Pise, N. (2023) 'CryptoScholarChain: Revolutionizing scholarship management framework with Blockchain technology', International Journal of Advanced Computer Science and Applications, 14(8).

Taherdoost, H. (2023) 'Smart contracts in Blockchain technology: A critical review', Information, 14(2), p. 117.

Xu, X., Weber, I. and Staples, M. (2019) Architecture for Blockchain applications.

Yan, J., Wang, K., Huang, Z., Zhang, R., Chen, N., Wei, L. and Zhou, Y. (2025) 'Educational institution financial management system based on enhanced smart contracts', in 2025 6th International Conference on Computer Engineering and Application (ICCEA), IEEE, pp. 1051–1055.

Yermack, D. (2024) 'Is Bitcoin a real currency? An economic appraisal', in Handbook of Digital Currency, pp. 29–40. Academic Press.

Zheng, X.R. and Lu, Y. (2022) 'Blockchain technology–recent research and future trend', Enterprise Information Systems, 16(12), p. 1939895.

Zheng, Z., Xie, S., Dai, H., Chen, X. and Wang, H. (2017, June) 'An overview of Blockchain technology: Architecture, consensus, and future trends', in 2017 IEEE International Congress on Big Data (BigData Congress), pp. 557–564. IEEE.

Human Cognition is a Tool in Warfare in the Cyber Domain

Niina Meriläinen
Faculty of Management and Business, Tampere University, Finland

Niina.merilainen@tuni.fi

Abstract: Warfare in cyber domains has evolved into a battleground for both defensive and offensive geopolitical objectives, exploiting cognitive vulnerabilities, credibility, and authority. Artificial intelligence (AI), information communication technologies (ICT), digital platforms and gaming are increasingly used to influence cognitive elements such as attitudes, biases, beliefs, perceptions, and psychological susceptibilities. Cognitive warfare is now an integral part of international relations, posing serious threats to individuals, critical infrastructure, and national security. Human cognition serves as a powerful instrument in this domain. As AI, ICT tools and digital platforms continue to advance, cyber warfare is becoming more complex, with diverse threat vectors and networks of adversarial actors. Tactics such as psychosensory warfare and white noise jamming are proving effective in cognitive operations, where fear and the loss of hope play a central role in digital battlefield strategies.

Keywords: AI, Cognitive warfare, Vulnerabilities, Internet, Digital platforms

1. Introduction

The cyber domain and various digital platforms have become arenas for both defensive and offensive geopolitical activities, including warfare. Cyber warfare exploits cognitive vulnerabilities such as biases, values, beliefs, credibility, and authority. In this process, artificial intelligence (AI), information and communication technologies (ICT), the Internet, social media, and gaming are actively employed. By targeting cognitive elements, such as biases, beliefs, attitudes, and understanding, cyber warfare seeks to shape individual and collective perceptions. Consequently, human behavior can be influenced, enabling adversaries to steer populations toward their strategic objectives. Human cognition is a central asset in cognitive warfare within the cyber domain. This form of warfare constitutes an integral part of modern conflict, posing serious threats to personal, organizational, military, and national security, and is conducted alongside kinetic warfare. The purpose of this article is to examine existing research and the thematic variations it presents, and to reflect on how future studies should approach this complex and evolving phenomenon through various threat vectors.

This theoretical paper draws on existing literature to explore the question:

How can human cognition be harnessed in cognitive warfare?

It provides a theoretical outlook and offers recommendations on why cognitive security should be considered a critical concern for:

- personal (tactical),
- municipality/organizational (operational) and
- state/military leadership (strategic) levels.

Warfare in the cyber domain involves manipulating the cognition of individuals, groups, or entire populations to gain strategic advantage over adversaries. It elevates human cognition into a critical dimension of modern warfare. The cyber domain provides an ideal environment for persistent and continuously evolving attacks. In this context, cognitive warfare integrates processes, knowledge, technologies, and cognitive capabilities related to threat awareness at personal, organizational, and state or military levels. Communication and cognitive attributes play a central role in these operations. While success in kinetic warfare can, at best, be measured in seconds, the strategic, operational, and tactical outcomes of cognitive warfare unfold gradually over time. Therefore, understanding human cognition—and how it can be harnessed as a tool alongside kinetic operations—is essential.

2. Related Works

2.1 Human Cognition and Warfare

Human cognition, while resilient, contains exploitable vulnerabilities. Cognitive warfare targets fundamental human needs, such as the need to belong and to feel loved and safe. It also leverages political and social silos, as well as mistrust toward leaders, legacy media, the military, police, politicians, and educators. People live within mental constructs (Lippmann, 1922). In the past, these were shaped by legacy media; today, they are influenced by actors from newer media, including influencers and keyboard warriors (Meriläinen, 2025;

Kalpokas, 2017). These influencers actively create, sustain, and exploit cognitive vulnerabilities for strategic gains in personal brand building as well as in cyber-based cognitive warfare. Cognitive warfare significantly overlaps with other forms of non-kinetic warfare. It serves as an umbrella concept encompassing psychological operations, information, political, cyber, and electronic warfare (Rhode, 2024). As Crawford (1999) notes, research on these topics is too important to be left solely to the military.

Human cognition, as it relates to warfare in the cyber domain, refers to the processes through which individuals learn, assess, and process information—encompassing perception, attention, language use, memory, and creative problem-solving (Ahmed et al., 2025). Cognitive warfare involves manipulating the cognition of individuals, groups, or entire populations—from everyday citizens to policymakers, military officers, and high-ranking officials. Its goal is to gain strategic advantage over adversaries by turning human cognition into a critical battleground. This form of warfare employs technologies such as AI, ICT, bots, advertisements, gaming, deepfakes, trolls, the Internet, and social media to alter the cognition of human targets—many of whom remain unaware of these manipulative efforts, as do those responsible for countering, mitigating, or managing their effects (Claverie & Du Cluzel, 2022; Meriläinen, 2025). Russia and China are recognized experts in leveraging cognitive elements in warfare (Fang, 2024). China has learned from Russia, with a key distinction: Russia disregards its international reputation, whereas China seeks to preserve it (Sobchuk, 2025). Other nations are following these two expert countries. Yet, Russia is a pioneer in psychological warfare, particularly in its methods of creating confusion and destabilization—not by offering alternative truths, but by undermining the very concept of truth itself (Sobchuk, 2025).

Beznosov (2025) explores how Russian psychosensory warfare targets human senses to influence emotions and, consequently, behavior. This strategy affects individuals and groups by manipulating sensory modalities—sight, touch, sound, taste, and smell—to provoke emotional responses. As Beznosov notes, such tactics foster a pervasive atmosphere of fear, with fear itself being the primary objective of psychosensory warfare. Continuous cognitive attacks by adversarial actors reinforce and sustain this climate, with the key being the absence of pauses between assaults. These attacks involve persistent messaging across multiple cyber and offline channels. To maximize impact, hostile actors conduct target group research to identify the most effective ways to amplify fear and hopelessness. Existing societal concerns—such as climate change, immigration, unemployment, and general uncertainty—are exploited as emotional triggers. Simultaneously, influencers are deployed to repeatedly disseminate these messages, ensuring constant exposure to a "wind of fear" (Meriläinen, 2025). Psychosensory warfare (Beznosov, 2025) operates in tandem with cognitive-level white noise jamming (Gatov, 2018), where external criticism is dismissed as Russophobia—echoing earlier Kremlin censorship of foreign broadcasts.

2.2 Cognitive Warfare and Societies Levels

The human mind is the battleground in cognitive warfare (Marjanović & Smiljanić, 2025). Not everything that happens in cyberspace stays there (Marsilii, 2023). In the cyber world, individuals navigate across multiple platforms and are influenced by a variety of actors and groups. There is no clear boundary between the analog and cyber spheres—they are deeply interconnected. Therefore, the connection between cognition and behavior must be understood in the context of warfare.

2.2.1 Personal level

Cognitive warfare targets the human mind by influencing values, biases, beliefs, and behavior, while eroding trust in governments, official actors, those around us, and the convictions people hold. Digital platforms, such as social media and Internet is used to undermine state interests, discredit institutions, and sow domestic discord (Prier, 2020; Meriläinen 2025). Individuals engage across various cyber arenas, receiving information, news, and even cognitive authority from alternative media sources, influencers, and actors outside traditional structures (Meriläinen, 2025). Many believe their media literacy is strong and assume immunity to tactics such as disinformation, AI- and ICT-generated content, deepfakes, and information operations (Meriläinen, 2024). ICT tools and disinformation play a central role in both defensive and offensive cognitive warfare (Lande & Danyk, 2025). In cyber domains, individuals are easily targeted (Meriläinen, 2025).

For instance, Russia's psychosensory warfare aims to influence foreign decision-making by supplying polluted information and fear (Beznosov, 2025). When disinformation becomes embedded in decision-making processes, Moscow achieves a key element of reflexive control (Beznosov, 2025). Even without direct policy impact, its spread through various digital platforms—such as social media—can create a permissive public opinion environment in which Russian narratives are perceived as factual (Kanet, 2024; Bykov, 2025;

Meriläinen, 2025). Kibeom (2025) argues that the proliferation of digital platforms has made disinformation a powerful tool. Due to cognitive biases and the effects of mis- and disinformation, countering fake news with accurate information is not always possible (Meriläinen, 2025; Lysenko et al., 2023). Manipulated information can reshape public perceptions (Lysenko et al., 2023). Moreover, ICT and global connectivity have enabled new and evolving threats from cyber domains—marked by unprecedented speed, scale, and unconventional tactics (Marsili, 2023).

Effective operations in cyber domains require strong situational awareness of events in both the physical and informational spheres, as well as an understanding of how these influence the cognitive domain (Ponangi et al., 2012). Cognitive warfare is a strategic approach to conflict that seeks to alter how target populations think and behave (Backes & Swab, 2019), leveraging content consumption and belief systems to shape individual thought processes. It is often deployed alongside both defensive and offensive kinetic warfare. Information operations are a central tool in cognitive warfare, designed to influence adversaries' cognition and achieve strategic objectives. These operations draw on a wide range of sources, including military doctrines, memes, online activism, disinformation, hashtags, historical narratives, humor, AI-generated fake accounts, imagery, algorithms, and bots (Meriläinen, 2024; Eleferenko, 2023; Weedon et al., 2017; Mozur et al., 2021).

The adversarial aspect of cyber-domain warfare frequently relies on government-funded networks and covert channels to disseminate disinformation, often through state actors or influencers (Meriläinen, 2025; Lysenko et al., 2023). For example, cognitive and information warfare, propaganda, and cyberattacks targeting Ukrainian infrastructure are designed to break the population's will to resist (Lysenko et al., 2023). As Kanet (2024) observes, cognitive warfare in cyber domains can alter perception to the point where disinformation is accepted as fact. The Russian military, for instance, employs AI and deepfakes in cyber-psychological and cyber-technical operations to achieve strategic, potentially war-winning effects. These influence operations aim to gradually undermine and weaken adversaries. Meanwhile, NATO lacks comprehensive policies to counter such tactics (Miron & Thornton, 2024). However, Bykov (2025) argues that NATO's cognitive security strategies and analysis of the informational environment can help counter Russia's cognitive warfare and related cyberattacks. This requires coordinated efforts across state institutions, civil society, the private sector, and international partnerships. Russia has effectively used information warfare to gain strategic advantages in conflicts, employing propaganda and psychological operations to divide enemy populations. Its leadership understands the power of information in shaping and manipulating audiences, and the rise of social media has made netizens increasingly vulnerable (Afridi, 2025). The individual human mind is connected to municipal, organizational, and military/state levels of society. When trust in societal structures and the actors within them begins to erode, and fear takes hold, the human mind becomes vulnerable to warfare. In such conditions, the mind itself becomes part of conflicts, sometimes even a soldier, a keyboard warrior.

2.2.2 Municipality and organizational levels

Warfare that leverages cognitive attributes as tools is a core element of modern international relations, posing serious threats to organizations and critical infrastructure, including banking, energy, transportation, national security, elections, economics, and politics (Musakhanov, 2023). Assets requiring protection include power plants, energy distribution networks, water treatment and supply facilities, transportation systems, radio and satellite communication networks, military communication channels, weapons systems, command and control centers, and surveillance and reconnaissance systems (Kumar & Nagar, 2024). Palletti et al. (2021) emphasize the interconnected nature of these systems and the cascading effects that cyberattacks can trigger. Attackers must understand how infrastructures are linked, which highlights the need for a global approach to strengthening cyber defense resilience in the face of increasingly frequent and sophisticated attacks on cyber domains—particularly those targeting cognition.

Executing cyberattacks against critical infrastructure requires the ability to modify malware and target specific industrial protocols—skills that are not widely available (Izycki & Vianna, 2021). Cognitive warfare encompasses a wide range of constantly evolving attacks targeting both physical systems and human cognition. Because these attacks can occur anywhere within cyber domains, they are not bound by geography and can be organized globally (Veljković, 2024). Targets include not only states and armed forces but also individuals, communities, and organizations. According to Veljković (2024), digital conflict is designed to harm adversaries, often with unintended consequences. It uses digital technologies to compromise, disrupt, or destroy cognition, information systems, networks, and infrastructure. Activities include espionage, sabotage, and propaganda, carried out by both state and non-state actors (Kumar & Nagar, 2024; Meriläinen, 2025; Veljković, 2024).

To carry out these attacks, adversaries must identify cognitive vulnerabilities. Hostile third parties may use computers for cybercrime and cyberwarfare, threatening the safety and security of individuals, their cognition, companies, and governments in ways that are difficult to counter using traditional methods (Dalton, 2017). The assets requiring protection remain the same: critical infrastructure and communication systems (Kumar & Nagar, 2024). Given the risks cyber warfare poses to both organizations and employees, it is essential to consider the cognitive impacts on staff, who may experience harassment and psychological strain from malicious attacks (Rafi & Imtiaz, 2023). Employees may feel disconnected from their workplace or colleagues. They may perceive decisions affecting them as being made elsewhere and feel unappreciated or distrusted by employers, politicians, and public officials—from healthcare to education and law enforcement. This alienation can lead individuals to identify with actors who promote distrust, finding common ground in the cyber domain. Everyday pressures may further erode their sense of belonging to their work, community, or society.

This dynamic is shaped by the constant priming and framing of cognitive vulnerabilities, values, biases, beliefs, emotions, and behaviors—factors linked to espionage, information warfare, cyberattacks, and influence operations. These processes threaten not only critical infrastructure but also societal cohesion. During information operations and warfare, cognitive attacks pose significant risks to global security, economic stability, and individual rights. Companies must adopt a holistic approach to address these threats (Sharma & Vandra, 2024), while hostile states exploit employee alienation. When individuals become disconnected from their organizations, their cognition can be manipulated in ways that turn them against their workplace—and even against their society and state. Notably, cognitive vulnerabilities, credibility, and authority are closely linked to social engineering. Social engineering involves manipulating human cognition to gain access to information and knowledge held by individuals during cyber-domain warfare. In this context, cognitive authority and perceived credibility are key tools for infiltrating targets—regardless of their social status or employment. Social engineering poses a serious security threat to infrastructure, users, data, and operations in cyberspace by exploiting vulnerabilities within cyber domains (Wang et al., 2020), spanning personal, organizational, state, and military levels.

2.2.3 State and military leadership levels

When states carry out cyberattacks, they may provoke counterattacks—highlighting both the defensive and offensive dimensions of warfare in cyber domains (Izycki & Vianna, 2021). Cognitive warfare in the cyber domain is invisible and insidious, creeping into its targets, and often, by the time its effects are recognized, it is too late for a state or military to respond effectively (Kibeom, 2025). It elevates traditional warfare by targeting how humans think, react, and make decisions, affecting civilians, decision-makers, and military personnel alike (Masakowski & Blatny, 2023). Scholars have examined whether military doctrines and counter-doctrines for cognitive warfare exist in countries such as Russia, China, Taiwan, and within NATO (Miron & Thornton, 2024; Bykov, 2025; Hung & Hung, 2022; Mozur et al., 2021; Reinhold & Reuter, 2022).

In offensive cognitive warfare, militaries may attempt to influence adversaries' perceptions of targets, troop movements, and future battle scenarios. In defensive operations, military leadership must train personnel to maintain trust in command structures, as adversaries aim to erode this trust through communication across both cyber and analog domains. Russia views ICT as a decisive factor in warfare, capable of influencing outcomes at multiple levels. The war in Ukraine has reshaped understandings of warfare on cyber domain, with superpower dynamics defining the nature and purpose of cyber operations. The relationship between ICT and cognitive methods continues to evolve (Kukkola, 2024). New techniques, simultaneous attacks, and asymmetric strategies are common in information warfare and related cyber operations (Hamilton et al., 2002). Cognitive attacks can bypass conventional defenses (Nikoula & McMahon, 2024). To effectively navigate cognitive warfare, military leaders must integrate traditional doctrine with modern digital tools and psychological strategies, ensuring resilience and strategic clarity in today's complex cyber battlespace. For example, resilience training should be incorporated into military education to counter cognitive warfare efforts (Rusu, 2025). Moreover, military leaders must prioritize the threat of disinformation and recognize that personnel often lack the skills and awareness needed to resist cognitive manipulation.

It is essential to recognize that military leaders and soldiers do not operate in a vacuum—they are influenced online like everyone else and are not immune to cognitive attacks. Even subtle cognitive manipulation can impair decision-making and challenge core values. To protect military personnel, leaders must implement proactive strategies, including training programs that enhance individual competencies in identifying and resisting disinformation (Cheatham et al., 2024). Operational, tactical, and strategic levels all face significant disadvantages in the cyber domain. As algorithmic manipulation and disinformation continue to spread,

greater emphasis must be placed on the strategic value of psychological operations and narrative control (Gombar, 2025). Some argue that national defense strategies must evolve to address not only technical vulnerabilities but also the social, psychological, and cognitive dimensions exploited by cognitive warfare. Both aspects are essential for protecting critical systems and the minds of individuals under attack (Meghraoui & Belkhamza, 2025).

3. Discussion and Conclusions

This theoretical study draws on existing literature and the themes it presents to explore the question: How can human cognition be harnessed in cognitive warfare within cyber domains? Human cognition is a powerful asset in cognitive warfare, which involves a wide range of continuous and ever-evolving attacks targeting both physical systems and the human mind. Because cognitive warfare can occur anywhere within both analog and cyber domains, it can be organized globally, removing geographical constraints and proximity requirements (Veljković, 2024). Furthermore, existing research indicates that cognitive warfare attacks are not limited to states and their armed forces, they also target individuals, communities, and organizations. Non-kinetic warfare can escalate into kinetic warfare and vice versa (Saressalo, 2025). Therefore, in future research we should focus on how these two forms of warfare interact, highlighting the harm caused by cognitive warfare—even when bloodshed is not immediately visible.

According to Veljković (2024) and Meriläinen (2025), conflict in digital environments is designed to harm adversaries, often with both intended and unintended consequences. It employs digital technologies to compromise, disrupt, or destroy cognition, information systems, networks, and infrastructure. Activities include various forms of AI, ICT, espionage, sabotage, and propaganda, conducted by both state and non-state actors (Kumar & Nagar, 2024; Meriläinen, 2025; Veljković, 2024). Personal (tactical), municipal/organizational (operational), and state/military leadership (strategic) levels must be viewed as a unified whole. We cannot protect just one part, as people form the foundation of all these societal levels. Moreover, individuals are influenced across all these levels through multiple vectors online and offline. These threat vectors permeate all levels and are constantly evolving, making it impossible to define any single, concrete, and lasting vector. This reality calls for continuous training across all levels of society we are seemingly unprepared to meet.

Individuals can be psychologically fragmented and repurposed as tools in cyber-domain warfare. Psychosensory warfare (Beznosov, 2025) and white noise jamming (Gatov, 2018) are effective instruments in this context. Fear and hopelessness play a central role in how cognitive warfare is operationalized and executed across digital arenas. Cyber resilience is essential for countering the effects of cognitive warfare, which exploits vulnerabilities in both digital systems and human cognition (Radu, 2025). Strategic, operational, and tactical levels must adopt vector-like procedures to defend against such threats. However, resilience is deeply tied to cognition—values, biases, and beliefs, making it difficult to build and sustain.

Because people have different understandings of reality, varying values, beliefs, and cognitive biases, how can we build awareness and protect ourselves, our organizations, armed forces, and states from hostile actors? For example, shortly after Russia's full-scale invasion of Ukraine, the Russian news agency TASS published disinformation claiming that "Zelensky hastily fled Kyiv" (TASS, 2022). This message was simultaneously spread on internet and social media by actors and accounts affiliated with Russia. Although legacy media outlets fact-checked and disproved the claim (DW, 2022; Reuters, 2022), many individuals no longer consume legacy media, leaving them vulnerable to distorted perceptions (Meriläinen, 2025). Cognitive warfare has continued since the invasion. Russian-affiliated social media accounts have blamed Zelensky for the lack of peace, portraying Russia as the party seeking resolution. Influencers use various cyber platforms, framings, and hashtags to push narratives suggesting Zelensky refused peace talks, labeling him a war criminal, and calling Ukraine supporters "people profiting from the war" and "retards" (Levinsz, 2025; Anna, 2025). These messages—spread by bots, deepfakes, and real accounts—are part of Russia's psychosensory and cognitive warfare strategy, aiming to instill fear and erode trust in national leadership, making surrender appear as the only viable option.

The core objectives of cognitive warfare in cyber domains are to instill and sustain fear, and to break trust: trust in one's environment, community, institutions, and leadership, including both civil, state and military authorities. The cyber domain offers an ideal platform for this, providing flexible and shifting threat vectors unconstrained by geography or time. By targeting values, biases, and beliefs—and by leveraging newer actors such as influencers, as well as the human need for hope and a vision of the future—adversaries can reshape cognition and influence behavior. Cyber domains thus serve as powerful arenas for cognitive warfare, where

human cognition is one of many tools that can be exploited at any time, cost-effectively and without restriction, to serve adversarial aims. As people increasingly use various digital platforms and are exposed to ICT- and AI-driven content, it is natural that warfare and influence operations have become a new frontier of war. Mastery in cognitive warfare lies in the fact that it is not always recognized as warfare, as its effects may be invisible or take a long time to manifest—unlike kinetic warfare, where outcomes can be seen within seconds. If the cyber environment and cognition are not acknowledged as arenas of warfare and properly prepared for, we risk enabling direct hostile operations against us.

Acknowledgements

I wish to thank the reviewer of this paper.

AI declaration: No AI tools were used in the creation of this paper.

Ethics declaration: No ethical clearance was required.

References

Afridi, M. (2025) "Deciphering Russian information warfare: Lessons from Georgia to Crimea and Ukraine", *ASSAJ*, Vol *3*, No. 1, pp 824-837.

Ahmed, S. S., Nisar, H., & Lo, P. K. (2025) "Introduction to human-machine interaction" In *Artificial Intelligence and Multimodal Signal Processing in Human-Machine Interaction*, Academic Press, pp 1-18.

Backes, O., & Swab, A. (2019) "Cognitive Warfare. The Russian Threat to Election Integrity in the Baltic States", Cambridge: Belfer Center for Science and International Affairs.

Beznosov D. S. (2025) "Social and psychological technologies of information warfare and means of counteracting them", Bulletin of the Saratov University, Series: Philosophy. Psychology, Vol 25, No. 2, pp 120-125.

Brittain-Hale, A. (2023). Clausewitzian Theory Of War in The Age of Cognitive Warfare. *The Defense Horizon Journal*, pp 1-4.

Bykov, S. (2025) "Challenges of Cognitive Warfare: The Ukrainian Case Study". *Thesis.*

Cheatham, M. J., Geyer, A. M., Nohle, P. A., & Vazquez, J. E. (2024) "Cognitive warfare: the fight for gray matter in the digital gray zone", *Joint Force Quarterly*, Vol 114, No. 2, pp 83-91.

Claverie, B. & Du Cluzel, F. (2022) ""Cognitive warfare": The advent of the concept of "cognitics" in the field of warfare", *Cognitive Warfare: the future of cognitive dominance*, pp 1-11.

Crawford, B. C. H. (1999) "Information warfare: Its application in military and civilian contexts", *The Information Society*, Vol 15, No. 4, pp 257-263.

Dalton, W. (2017) "The silicon hat hacker: using reinforcement learning in hybrid warfare. Conference: Hybrid Threats and Asymmetric Warfare: What to do?", Swedish Defence University (SEDU) – Stockholm.

DW (2022) "Zelenskyy rejects rumors he has fled Kyiv — as it happened" (online) https://www.dw.com/en/ukraine-zelenskyy-rejects-rumors-he-has-fled-country-as-it-happened/a-60908297

Eleferenko, A. (2023) "How can digital diplomacy reconcile Russia and the West?", *International Policy Review,* Vol. 4, No. 1, pp 1-20.

Fang, K. (2024) "Wangbao (Cyberbullying) and Jubao (Reporting): Collaborative State-Society Online Influence Operations in China" *Journal of Online Trust and Safety*, *2*(3).

Gatov, V. (2018) How to Talk with Russia. Public diplomacy for the 21st century, [online] https://www.lse.ac.uk/iga/assets/documents/research-and-publications/How-to-talk-with-Russia-Public-Diplomacy-for-the-21st-Century-3.pdf

Gray, G. P. (2025) "The Political Expediency of Counter-Disinformation: An Analysis of the Response to Russian Information Warfare", (online) https://tsuda.repo.nii.ac.jp/record/2000108/files/%E5%9B%BD%E9%9A%9B%E9%96%A2%E4%BF%82%E5%AD%A6%E7%A0%94%E7%A9%B651_01-16_Gray.pdf

Gombar, M. (2025) "Algorithmic Manipulation and Information Science: Media Theories and Cognitive Warfare in Strategic Communication", *European Journal of Communication and Media Studies*, Vol *4, No.* 2, pp 1-11.

Hamilton, S. N., Miller, W. L., Ott, A., & Saydjari, O. S. (2002) "Challenges in applying game theory to the domain of information warfare", In Information Survivability Workshop (ISW).

Hung, T. C., & Hung, T. W. (2022) "How China's cognitive warfare works: a frontline perspective of Taiwan's anti-disinformation wars", *Journal of Global Security Studies*, Vol *7*, No. 4, pp 1-18.

Izycki, E., & Vianna, E. W. (2021) "Critical infrastructure: A battlefield for cyber warfare". In *ICCWS 2021 16th international conference on cyber warfare and security*, pp 454-462.

Kalpokas, I. (2017) "Information warfare on social media: A brand management perspective", *Baltic Journal of Law & Politics*, Vol 10, No. 1, pp 35-62.

Kanet, R. E. (2024) "Moscow and the World: From Soviet Active Measures to Russian Information Warfare", *Applied Cybersecurity & Internet Governance*, Vol 3, No.1, pp 34-57.

Kibeom, K. (2025) "A Study on the Concept of Cyber Cognitive Warfare and Case Study Methodology From a Psychological Perspective", In *European Conference on Cyber Warfare and Security*, pp 241-249.

Kukkola, J. (2024) “Suvereenit hiekkamadot: Venäjän kybertoiminta osana valtioiden välistä kamppailua 2000-luvulla” (online) https://www.doria.fi/handle/10024/188804

Kumar, S., & Nagar, G. (2024) ”Threat Modeling for Cyber Warfare Against Less Cyber-Dependent Adversaries”, In *European Conference on Cyber Warfare and Security*, Vol 23, No. 1, pp 257-264.

Lande, D., & Danyk, Y. (2025) ”Competitive Artificial Intelligence in Information and Cyber Warfare”, (online) *SSRN 5084698*.

Lippmann, W. (1922) Public Opinion, New York, USA. Macmillan.

Lysenko, S., Marukhovskyi, O., Krap, A., Illiuschenko, S., & Pochapska, O. (2023) “The Analysis of World Information Warfare and Information Security in the Context of the Russian-Ukrainian War”, *Studies in media and communication*, Vol 11, No. 7, pp 150-158.

Marjanović, A., & Smiljanić, D. (2025) “Cognitive warfare–the human mind as the new battlefield”, In *Proceedings of the Defense and Security Conference,* pp. 84-114.

Marsili, M. (2023) “Guerre à la Carte: Cyber, information, cognitive warfare and the metaverse“ *Applied Cybersecurity & Internet Governance*, Vol 2, No. 1, pp 1-11.

Meghraoui, L., & Belkhamza, Z. (2025) “Cognitive Warfare and Cybersecurity: Strategic Implications for Global Security”, In *Proceedings of the 19th International Conference on Cyber Warfare and Security*”, pp 257-264.

Meriläinen, N. (2025) “Influencers as Tools in Hybrid Operations Online”, Proceedings of the 20th International Conference on Cyber Warfare and Security, Vol. 20, No. 1, pp 256-272.

Meriläinen, N. (2024) “The possible role of digital platforms in information operations”, Proceedings of the 11th European Conference on Social Media - ECSM 2024, Vol 11, No. 1, pp 137–143

Meriläinen, N., Hiljanen, M., & Rautiainen, M. (2023) Archetypes of youth as vectors in power relations From praises to information operations. Frontiers in Political Science, 5, 1228838.

Miron, M. & Thornton, R. (2024) “The Use of Cyber Tools by the Russian Military: Lessons from the War against Ukraine and a Warning for NATO?”, *Applied Cybersecurity & Internet Governance*, Vol *3*.

Mozur, P., Zhong. R., Krolik, A. Aufrichtig, A. & Nailah Morgan N. (2021) INSIDE A CHINESE PROPAGANDA CAMPAIGN, [online] The New York Times, https://www.nytimes.com/interactive/2021/12/13/technology/china-propaganda-youtube-influencers.html

Musakhanov, D. (2023) “The international consequences of cyber warfare: A study of the" Stuxnet" case”, *Acta of Turin Polytechnic University in Tashkent*, Vol *13, No.*3, pp 47-50.

Nikoula, D., & McMahon, D. (2024). Cognitive warfare: Securing hearts and minds. *Information Integrity Lab. JULY*. pp 1-16.

Palleti, V. R., Adepu, S., Mishra, V. K., & Mathur, A. (2021) ”Cascading effects of cyber-attacks on interconnected critical infrastructure”, *Cybersecurity*, Vol *4*, pp 1-19.

Ponangi, P., Kidambi, P., Rao, D., Fendley, M., Haas, M., & Narayanan, S. (2012) “On the Offense: Using Cyber Weapons to Influence Cognitive Behavior”, *International Journal of Cyber Society and Education,* Vol 5, No. 2, pp 127-150.

Prier, J. (2020) “Commanding the trend: Social media as information warfare”, In *Information warfare in the age of cyber conflict*, Routledge, pp 88-113.

Rafi, S., & Imtiaz, N. (2023) “Cyberwar: Its Psychological Impact on Employees and Consequences for Organizations”, In *Handbook of Research on War Policies, Strategies, and Cyber Wars,* IGI Global, pp 108-127.

Radu, R. (2025) “Building Cyber Resilience to Face the Challenges of Cognitive Warfare”, In *European Conference on Cyber Warfare and Security*, pp 803-810.

Reinhold, T., & Reuter, C. (2022) Artificial Intelligence and The Future of Warfare, The USA China and Strategic Stability. Manchester. Manchester University Press.

Reuters (2022) “Fact Check: President Volodymyr Zelenskiy still in Ukraine as of Oct. 25, contrary to online claims” (online) https://www.reuters.com/article/fact-check/president-volodymyr-zelenskiy-still-in-ukraine-as-of-oct-25-contrary-to-online-idUSL1N31Q1NU/

Rhode, S. (2024) “A Systematic Review of Psychological Warfare, Cognitive Warfare, Associated Terminology and Techniques” Master Thesis. der Helmut-Schmidt-Universität / Universität der Bundeswehr Hamburg.

Rusu, A. (2025) “Military Ethics in the Light of Cognitive Warfare”, *Redefining Community in Intercultural Context*, Vol *12*, No. 1, pp 45-50.

Saressalo, T. (2025) “Information influencing in wars and conflicts in the early 21st century”, National Defence University Series 1, Research Publications, 67.

Sobchuk, M. (2025) ”To understand the nature of modern Chinese influence operations, study Russia first”, Cyfluence Research Center, pp 1-7.

Tass (2022) “Zelensky hastily fled Kiev, Russian State Duma Speaker claims” (online) https://tass.com/politics/1411855

Weedon, J., Nuland, W. and Stamos, A. (2017) Information operations and Facebook. [Online] https://fbnewsroomus. files. wordpress. com/2017/04/facebook-and-information-operations-v1. pd

Veljković, S. (2024) “Possibility of applying the rules of international humanitarian law to cyber warfare”, *Pravo teorija i praksa*, Vol 41, No.3, pp 17-28.

Wang, Z., Sun, L., & Zhu, H. (2020). Defining social engineering in cybersecurity. *IEEe Access*, *8*, 85094-85115.

A Hybrid Machine Learning Approach for Red Team Log Analysis

Nickolas Mohr, Alan Shaffer, Gurminder Singh and Armon Barton
Naval Postgraduate School, Monterey, USA

nickolas.mohr@nps.edu
alan.shaffer@nps.edu
gsingh@nps.edu
armon.barton@nps.edu

Abstract: Red teaming is a common cybersecurity practice that simulates real-world adversarial cyber operations on defended systems to identify vulnerabilities. Current red team tools often have limited logging capabilities, resulting in insufficient analysis that prevents red teams from receiving real-time feedback and insights after operations. The application of machine learning for automating the analysis of red team operations is severely constrained by the scarcity of labeled, real-world log data. This research addresses this challenge by exploring the potential of using synthetic data to train attack-detection models for Cobalt Strike logs. We systematically evaluate three different training approaches for analyzing Cobalt Strike operational logs: synthetic-only, real-world-only, and a hybrid approach that combines both data types. Our methodology employs a comprehensive feature engineering pipeline that includes both programmatic log generation for creating large-scale structured data and large language model techniques for introducing variety and edge cases. We transform each log file into a high-dimensional vector that includes event types, command verbs, temporal activity patterns, and mappings to the MITRE ATT&CK knowledge base. Random Forest classification models are trained using this feature set to distinguish between successful and failed attack scenarios. By rigorously testing each training approach against a manually labeled ground-truth set of 112 authentic Cobalt Strike logs, we quantify the performance and limitations of each strategy. Our main contribution is demonstrating that a hybrid training strategy achieves 94% accuracy, greatly surpassing synthetic-only models (56%) and real-world-only models (79%). This combined approach effectively addresses both the domain gap in synthetic data and the data scarcity in small, real-world datasets. The hybrid model learns attack diversity from over 30,000 synthetic scenarios while grounding understanding in the authentic structural patterns of real logs, providing a 15 percentage-point improvement over real-data-only approaches. This research offers a practical framework for enhancing limited real-world cybersecurity datasets by strategically integrating synthetic data, enabling immediate use in Department of Defense red team operations and wider cybersecurity machine learning applications.

Keywords: Red team, Log analysis, Machine learning, Vulnerability assessment, Large language model

1. Introduction

In the modern cybersecurity arena, an organization is constantly under threat from malicious actors seeking to exploit its vulnerabilities. This digital arena is like a battlefield where attacks are not just possible but unavoidable. Therefore, it is critical for an organization like the U.S. military, which manages a huge amount of sensitive data, to prepare for potential incidents in cyberspace.

Red team operations are a crucial component of a proactive cybersecurity plan, as they simulate real-world attacks. Red teams help organizations identify weaknesses in their systems, processes, and staff before malicious actors can find and exploit them. The importance of red teaming cannot be overstated, especially in the realm of national security. As reviewed in the 2024 Department of Defense (DoD) Cyber Assessment Program report, the offensive capabilities of potential adversaries are escalating; many cyber defense and operations missions remain vulnerable to offensive cyber capabilities of potential adversaries (Director, Operational Test and Evaluation, 2024).

Red teaming has its roots in military strategy, where commanders tested battle plans by designating a group to play the role of the adversary. In the context of cybersecurity, the National Institute of Standards and Technology defines a red team as "a group of people authorized and organized to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture" (National Institute for Standards and Technology, 2025).

While the importance of red teaming is widely recognized, little attention has been given to transforming their activity data, namely, log data captured in tools such as Cobalt Strike, into structured insights for analysis. These logs provide detailed information about operator actions, tools used, and mission outcomes; however, logs are often noisy and inconsistent. Another impediment is the significant lack of high-quality, labeled red team log data for analysis.

This research has addressed these problems using a hybrid synthetic data approach. We developed a programmatic log generator that produces labeled data at scale with consistent structure and accuracy. Large language model methods introduce variety and rare cases, but struggle with longer sequences and require careful quality control. Combining the LLM-generated logs with the programmatically generated logs establishes a training foundation that supports the development of accurate models and exposes them to unusual patterns. This research has demonstrated that the automated analysis of red team log data is feasible and that linking red team exercise data with machine learning can enhance insights into red team effectiveness.

2. Related Work

Red team tools generally fall into three categories: tools with advanced capabilities but devoid of logging features, those with limited or standardized logging capabilities, and those that provide extensive logging with some form of log analysis functionality.

Tools such as the Social Engineering Toolkit (SET) and Nmap are prime examples. These are open-source tools designed for easy use, but their focus on functionality comes at the expense of their inability to produce comprehensive user logs (Lyon, 1997; Kennedy, 2010).

The Metasploit Framework (MSF) is a popular open-source application for penetration testing, providing an extensive toolkit for employing and testing various payloads against targets. MSF offers a wide array of logging features, including logging a session, logging a database attack, and logging from within the MSF console (Moore, 2003). However, MSF cannot extract meaningful insights from its logs. It lacks data visualization, pattern matching, and automation features to help users identify successful actions, failures, or anomalies during penetration testing.

Similarly, PowerShell Empire offers red team operators a wide range of actions to take on a target host that has already been compromised. PowerShell Empire can log user actions, including agent check-ups, tasking, and executions, within the compromised system (Schroeder and Ross, 2015). The logs themselves offer a highly detailed record of the actions taken post-exploitation; however, like MSF, PowerShell Empire does not provide any pattern recognition or anomaly detection capabilities for logged data.

Log data is diverse, unstructured, and enormous in scale. It is generated by servers, devices, applications, operating systems, and networks, with each producing unique formats that can further complicate the data aggregation process (Raval and Verma, 2017)**.** Developers have created log management tools such as Splunk (Baum, Das, and Swan, 2003), Elastic Stack (Banon, 2010), and Graylog (Koopmann, 2010) to manage log data; however, these tools often struggle with processing diverse and unstructured logs (Novikov, 2017).

The Elastic Stack is an open-source solution that appeals to organizations for log storage, management, and analysis. However, the appeal quickly diminishes when data volume and complexity increase, requiring significant processing power to maintain performance (Banon, 2010). On the enterprise side, Splunk has established itself as the industry leader, particularly excelling in security analytics and IT operations. It enables users to view real-time analytics and insights from ingested log data. However, current licensing fees for Splunk can amount to $20,000 annually, not including the implementation of specialized analysis tools (Baum, Das, and Swan, 2003).

High false positive rates in log analysis tools lead to "alert fatigue," where analysts overlook genuine threats due to overwhelming volumes of false alerts (Tariq *et al.*, 2025). Advanced features, such as machine learning and anomaly detection, can help reduce false positives; however, not all tools have these capabilities built in. Even when available, such features often require significant additional setup and domain expertise.

Techniques such as supervised learning (decision trees, support vector machines), unsupervised learning (clustering), and reinforcement learning have proven effective in various cybersecurity applications, including anomaly detection and intrusion prevention (Buczak and Guven, 2015). The application of machine learning techniques to cybersecurity log analysis faces a fundamental challenge: the scarcity of high-quality, labelled training datasets. This problem is particularly acute in specialized domains, such as red team operations, where data is often classified or operationally sensitive. Traditional supervised learning approaches require substantial datasets of labelled examples to achieve reliable performance, creating a significant barrier to developing automated analysis capabilities for red team exercises.

3. Methodology

This section presents the hybrid synthetic data approach developed for automated red team log analysis. We describe the three data generation methods, feature engineering pipeline, and machine learning implementation used to classify operational outcomes from Cobalt Strike logs.

3.1 Overview and Design Goals

The core motivation for this research stems from the enormous challenges inherent in manually processing Cobalt Strike logs generated during red team operations. These logs, although rich in valuable information, are often dense, unstructured, and inconsistent in format, making manual review extremely time-consuming and prone to errors for operators and analysts. Traditional machine learning approaches face a fundamental barrier: the scarcity of labeled training data due to classification restrictions and operational security concerns.

This research addresses these challenges through a hybrid synthetic data approach that combines three distinct data sources: programmatically generated logs for scalability and consistency, LLM-based generation for diversity and edge cases, and authentic operational logs for grounding in real-world patterns. The methodology focuses specifically on Cobalt Strike logs, developing a comprehensive feature engineering pipeline that transforms raw log entries into structured, machine learning-ready datasets.

The primary scope of this work is to parse and analyze log files generated by the Cobalt Strike command and control framework. Our analyses focused on Cobalt Strike due to its widespread use within the U.S. military, and its well-documented format structure. The pipeline parses logs, links events to MITRE ATT&CK techniques, and uses Random Forests to classify operation outcomes, comparing synthetic-only, real-world-only, and hybrid training approaches to demonstrate the effectiveness of the combined methodology.

3.2 Data Generation Approaches

To address the fundamental challenge of data scarcity in red team log analysis, this research employs a three-pronged approach to data generation. Each approach contributes distinct advantages: programmatic generation provides scalability and structural consistency, LLM-based methods introduce tactical diversity and edge cases, and real-world logs ground the models in authentic operational patterns. The combination of these approaches enables comprehensive model training while maintaining the ability to validate performance against genuine red team data.

3.2.1 Programmatic log generation

A custom Python program named log_gen.py was created to generate synthetic logs that resemble those produced by Cobalt Strike. The generator utilizes randomized parameters, including beacon sessions, beacon IDs, IP addresses, and commands, to enhance realism and variety. Each log entry clearly shows whether it pertains to beacon initialization, tasking, or output, making precise and automated tracking easier during analysis. The log_gen.py file can be configured to generate dozens, hundreds, or even thousands of synthetic logs, each with a definitive ground-truth label indicating whether the simulated attack was successful.

The log sessions generated by log_gen.py use randomization and generate various MITRE ATT&CK techniques, including lateral movement, credential theft, and privilege escalation. Sequenced commands and their corresponding outputs identify successful attack chains, while failure chains are characterized by syntax errors, permission denials, or "access-denied" messages. Each synthetic log session is configured to randomly represent varying operator identities (such as "trainee1," "PenTester," "AlphaOp"), infrastructure components (including "DOMAIN," "ACME"), hostnames (e.g., "DC01," "Laptop-User1"), services (such as HTTPS, DNS, SMB), and ports (for example, 443, 80, 53). This approach allows the batch runner to efficiently automate the generation of thousands of log entries with balanced labeling.

3.2.2 Large language model-based generation

While the programmatic generator provided a scalable foundation of structurally consistent data, a second method was employed to enhance the dataset's diversity and introduce less predictable attack narratives. We generated some 150 additional logs using OpenAI's ChatGPT-4, as well as Claude and Gemini 2.5, following a controlled implementation process designed to ensure the quality and realism of the synthetic logs. The implementation of this LLM-powered generation began with detailed prompt engineering to guide the models' output. Rather than using generic requests, prompts were carefully crafted to provide the models with specific context, constraints, and a desired format based on the sanitized, real-world log examples in our possession.

3.2.3 Real world ground truth dataset

The foundation for all model validation and the anchor for our hybrid training approach was a set of 112 authentic Cobalt Strike logs derived from a real red team exercise. This dataset, referred to as FullLogAnalyzed, served as the definitive ground-truth for this research. Each log in this batch was manually analysed to determine the ultimate operational outcome of the red team operator's actions. This manual analysis generated a labelled dataset comprising 58 logs representing failed attack scenarios and 54 logs representing successful scenarios. While limited in size compared to the synthetic datasets, this batch was invaluable. It provided the authentic structural patterns, rhythms, and beacon life cycle events that our initial analysis had proven were missing from the synthetically generated logs, making it an essential component for both model validation and successfully bridging the synthetic-to-real gap.

3.3 Feature Engineering

The conversion of raw log data into predictive analytics was facilitated by a range of Python modules, each with a distinct function in the workflow, as illustrated in Figure 1. The analytical pipeline begins with preprocessor.py, which plays a crucial role in preprocessing and feature discovery. This first step is performed once to scan the entire collection of available log files, including both synthetic and real-world variations. Its goal is to identify and gather all unique features within the complete dataset. Features include various event types (e.g., beacon_checkin), operator names, effective command verbs (e.g., getsystem), and MITRE ATT&CK techniques. The output of this script is a JSON file called feature_metadata.json. This file acts as a master blueprint and comprehensive list of all potential features across the dataset. By establishing this data, it ensures that every feature vector generated in subsequent steps maintains a consistent structure.

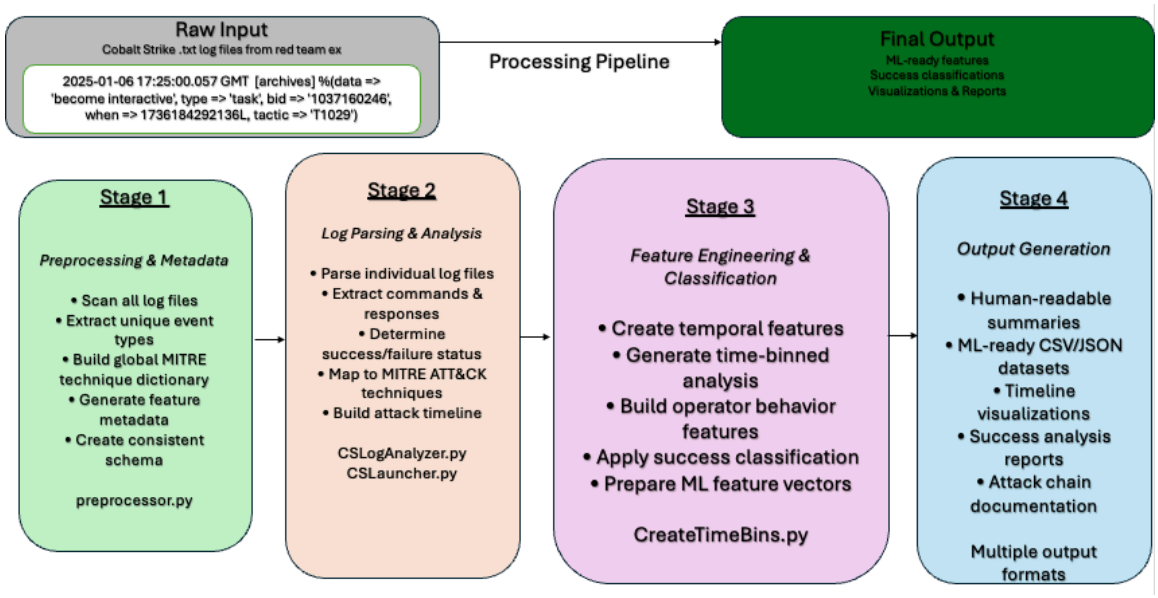


Figure 1: High-Level Architecture of the Cobalt Strike Log Analysis Tool

The log parser's primary role is to accurately calculate the number of these known features present in a specific log. It processes raw log entries and transforms them into structured data for analysis and ML feature extraction. The parser is specifically designed to handle Cobalt Strike logs, processing several key elements from each log entry: timestamps extracted in format "2025-07-01 03:14:07," event types categorized as "beacon_taskout" or "beacon_output," specific commands like "jump 10.0.2.8" or "hashdump," beacon IDs as unique identifiers for each session, operator information, source and target IPs extracted from network operations, session context derived from beacon metadata and prior activity, and success or failure labels determined from filename.

A key part of the data preparation process involves enrichment and labeling through MITRE ATT&CK technique mapping. The system incorporates MITRE ATT&CK technique labels through the Mitre_Mapping.py file, which contains a Python dictionary (MITRE_LOOKUP) that associates Cobalt Strike command verbs with corresponding MITRE ATT&CK techniques by their unique IDs and names. This structure enables a single Cobalt Strike verb to map to multiple techniques, providing standardized categorization of observed adversarial actions.

The final ML feature vector is structured to include the source log filename, one-hot encoded event types and MITRE techniques, log-level operator features such as binned start hour and total duration, and one-hot encoded time bin activity across 288 five-minute intervals over a 24-hour cycle. In this research, a total of 781 unique features were identified, creating comprehensive feature vectors suitable for machine learning analysis.

3.4 Machine Learning Implementation

With the data processing pipeline established, the final stage of implementation focused on the ML components. This involved two key phases: generating a comprehensive, ML-ready feature vector from the structured log data and selecting and implementing the algorithm used for training and validation.

For this research, the random forest classifier was selected as the primary ML algorithm. We chose this data classification model for several key reasons. It consistently performs well on structured data, such as the sets generated in our pipeline. It is also robust against overfitting, which is a significant risk when working with a high-dimensional feature space like our 781 features. Most importantly, a Random Forest offers a high level of interpretability through its feature importance scores, which highlight the features the model finds most predictive.

To conduct comprehensive testing, three distinct datasets were created and used in various combinations. The first was a Real-World Ground-Truth Dataset (FullLogAnalyzed), an authentic collection of 112 manually labeled logs from real-world red team operations. The second was an LLM-generated synthetic dataset (FullLogGpt), containing 123 logs with different and creative attack scenarios produced by LLMs. The final and largest source was a Programmatic Synthetic Dataset (LargeScaleTestData), which included over 30,000 logs generated by the Python script to ensure both massive scale and consistent scenarios.

To ensure a fair comparison between models, a single, consistent holdout set was used for final validation; the Real-World Ground-Truth Dataset was reserved for this purpose. Regardless of what data a model was trained on (synthetic, real, or a hybrid), its ultimate performance was judged based on its accuracy on real-world data, providing an unbiased comparison for every test case.

4. Results and Analysis

Our testing compared the performance of seven different models to answer three key questions. First, how well does a model trained on synthetic data perform on real data? Second, what is the best performance we can achieve using only a small set of real data? Finally, can a 'hybrid' model, trained on a blend of both data types, overcome the limitations of the other approaches to achieve the highest accuracy?

A central challenge in this research was the "synthetic-to-real" domain gap, which refers to the difference between the data patterns of a synthetic training dataset (the source) and a real-world dataset (the target). An ML model is said to generalize if it can accurately make predictions on new, unseen data that is different from the data it was trained on. If the model only memorizes its training data without learning any of the underlying issues, it fails to generalize.

Our initial tests were designed to measure this effect. The first model was trained exclusively on the LLM-generated synthetic data. While it achieved a near-perfect accuracy on its own test data (97% accuracy), its performance on the real-world holdout set collapsed to 48% accuracy. The model failed to generalize, as it learned and applied incorrect rules from the synthetic data to the real-world data. It was unable to identify a single failed attack in the real-world dataset correctly.

A second test model trained solely on synthetic data was developed using a much larger programmatic dataset. Its performance on real-world data slightly improved to 56%, but it still failed to generalize reliably, performing marginally better than a coin flip. A final model trained on a combination of both synthetic sources also underperformed, reaching only 48% accuracy. These experiments clearly show a significant domain gap, indicating that even a large and diverse collection of purely synthetic logs lacks the necessary patterns for a model to learn the correct rules and effectively generalize to real-world operational data.

To establish a performance baseline using only authentic data, a model was trained exclusively on the 112 manually labeled real-world logs. This model performed significantly better than the synthetic-only test cases, achieving an accuracy of 79%. This result is significant because it demonstrates that real-world logs contain consistent, learnable patterns that are fundamentally different from those in synthetic datasets. Additionally,

it highlights the issue of data scarcity. While 79% is a much better performance than seen with the synthetic models, the small number of training data ultimately limits the model's performance.

The hybrid approach proved to be a great success, significantly surpassing the baseline models. The Programmatic-Hybrid model achieved a peak performance of 94% accuracy. Its high precision (0.93-0.95) and recall (0.93-0.94) across both classes confirm that it learned to generalize effectively. This demonstrates that combining a large-scale, consistent synthetic foundation with a small authentic dataset is the optimal strategy for building a robust and reliable model.

4.1 Performance Summary

The comparison of the seven training models revealed interesting results and a clear order of performance. The experiments not only provided insights into the specific challenges of the synthetic-to-real domain gap and data scarcity but also confirmed the hybrid data approach as the best solution. As illustrated on Table 1, the Programmatic-Hybrid model, which combined large-scale synthetic data with a high-fidelity real-world dataset, emerged as the optimum model.

Models trained solely on real data reach a baseline accuracy of 79%, while integrating synthetic and real logs boosts accuracy to 94%. This demonstrates that programmatic generation provides scalability, LLMs introduce diversity, and real logs ground the model in authentic patterns. The approach also paves the way for future work in operator profiling through multi-class classification, as well as exploring other generative models like autoencoders and generative adversarial networks and applying the hybrid method to other fields with limited data.

The primary objective of this research was to develop and validate an effective ML pipeline for classifying the outcomes of red team operations based on log data. The comparison of three different training strategies revealed a performance hierarchy, offering clear insights into the challenges and solutions involved in using synthetic data for this task. The experiments began by establishing a baseline for a model trained exclusively on synthetic data, which achieved near-perfect accuracy (97%) on its synthetic test data but performance significantly declined when evaluated against a holdout set of 112 real-world logs, reaching only 56% accuracy. This confirmed the existence of a synthetic-to-real domain gap.

To establish a second baseline, a "real-only" model was trained exclusively on the 112 manually labeled real-world logs, achieving a respectable 79% accuracy. This result underscored the critical importance of authentic data patterns while simultaneously highlighting a performance ceiling imposed by data scarcity. The final experiment validated the "Hybrid Model" as the best solution, achieving an outstanding 94% accuracy on the same pure real-world validation set. This result represents a 38 percentage-point improvement over the best synthetic-only model and a 15 percentage-point improvement over the real-only baseline, demonstrating that the hybrid approach effectively addresses both the domain gap and data scarcity problems simultaneously.

Table 1: Comparative Accuracy of Training Methodologies on the Real-World Holdout Dataset

Training Method	Training Data	Data Ratio	Accuracy on Real-World Data
LLM-Synthetic Only	LLM-Generated	100% LLM	48%
Synthetic-Blend	LLM-Generated + Programmatic	1% LLM / 99% Programmatic	48%
Programmatic-Synthetic Only	Programmatic	100% Programmatic	56%
Real-World Only	Real-World	100% Real-World	79%
LLM-Hybrid	LLM-Generated + Real-World	57% LLM / 43 % Real	93%
Ultimate-Hybrid	Programmatic + LLM Generated + Real-World	98 % programmatic / 1 % LLM / 1 % Real World	93%
Programmatic-Hybrid	Programmatic + Real-World	99 % Programmatic / 1% Real World	94%

5. Discussion and Future Work

This research has demonstrated how the analysis of red team operations can be enhanced by creating and testing an ML pipeline that addresses log data scarcity, a key issue in the cyber domain. The primary challenge in this work was to determine if a model's predictive accuracy could be significantly enhanced by incorporating a substantial amount of synthetic data into a small, high-quality anchor dataset. By comparing seven different data set models, we developed an effective plan for overcoming the problem of data scarcity.

This research demonstrated that a small set of real-world log data (which alone can train a model to only 79% baseline accuracy) can be significantly improved by combining this data with a large synthetic dataset, to form a new Programmatic-Hybrid data model. When tested against the complete set of 112 real-world logs, this Hybrid Model achieved an impressive 94% accuracy rate. This demonstrates that real-world data instructs the model on what to look for, while the large synthetic dataset enables it to learn how to generalize that knowledge across thousands of variations.

This discovery shows that developing high-performing ML models does not require a large (and often hard-to-collect) set of real-world data samples. Instead, by using an approach that combines quality synthetic data with real-world operational log data, it becomes possible to create more accurate and effective data analysis models.

5.1 Implications for the Department of Defense

The insights gained from this research were transformed into actionable intelligence by developing an ML model that learns the signatures of a successful attack. This approach goes beyond simple log parsing and advances into predictive analysis. The model identifies actions among features, revealing patterns that may indicate a successful operation, such as a high frequency of specific data exfiltration techniques combined with a low frequency of error events. Once trained on the hybrid dataset, this model becomes a reusable tool that can be applied to future, unseen real-world logs to deliver data insights.

5.2 Future Research Direction

The primary approach of this thesis was to address log data scarcity by utilizing large-scale synthetic data anchored to a small, high-quality set of real-world data, which was shown to be highly adaptable. Future research should focus on applying this framework to new analysis areas and exploring more advanced generative techniques.

A significant impact of this research will be its potential application beyond cybersecurity. Many critical ML fields, particularly in the medical domain, are limited by the scarcity of large, labeled datasets due to privacy concerns and restrictions. A key area for future research is to perform experiments to validate this hybrid methodology in these other areas.

The model in this thesis was trained as a binary classifier to distinguish between success and failure. A logical next step is to develop a multi-class classifier for operator profiling. By training the model on logs labeled by individual operators, the pipeline could learn to identify each user's unique "signature" (such as their preferred tools, common errors, and operational tempo).

An interesting next step would be to adapt the log parser and log generation programs to support other common red team tools and frameworks, such as Metasploit and PowerShell Empire. This would confirm the validity of the hybrid training method across various log formats and help develop a more universal tool for automated red team analysis.

6. Conclusions

This research has demonstrated that the automated analysis of red team log data is feasible and that linking red team exercise data with machine learning can enhance insights into red team effectiveness. The hybrid synthetic data approach developed addresses the fundamental challenge of data scarcity in cybersecurity machine learning applications while maintaining the ability to achieve high performance on real-world operational data.

The key findings demonstrate that models trained solely on real data reach a baseline accuracy of 79%, while integrating synthetic and real logs boosts accuracy to 94%. This 15 percentage-point improvement represents a significant advancement in automated red team log analysis capabilities. The research validates that programmatic generation provides scalability, LLMs introduce diversity, and real logs ground the model in

authentic patterns, creating a comprehensive training methodology that overcomes both domain gap and data scarcity limitations.

The approach provides a practical framework for enhancing limited real-world cybersecurity datasets by strategically integrating synthetic data, enabling immediate use in Department of Defense red team operations and wider cybersecurity machine learning applications. By demonstrating that high-performing ML models can be developed without requiring large collections of classified operational data, this research opens new possibilities for advancing cybersecurity analytics while maintaining operational security requirements.

Ethics Declaration: This research involved analyzing cybersecurity log data from red team exercises. Ethical clearance was not required for this research as it utilized anonymized log files from existing Department of Defense training exercises with all personally identifiable information removed.

AI Declaration: The authors used generative AI to support multiple aspects of this research, including code development, idea generation, synthetic data creation, and writing improvement. Claude 3.5 Sonnet, ChatGPT 4, and Gemini 2.5 were used for several purposes: (1) to help debug and generate code for the log analysis tool, (2) to brainstorm ideas and concepts during the initial research process, and (3) to help generate synthetic Cobalt Strike logs that were used to train ML models. Grammarly was also used to proofread drafts and improve grammar, punctuation, and sentence structure during the writing process.

For code development, AI tools were provided with code functions and error messages, with suggested revisions implemented and tested using an iterative process. For idea generation, conversational interactions were used to explore research concepts and refine the approach. For synthetic log generation, AI was used to help create realistic training data that mimicked authentic Cobalt Strike log formats and scenarios. For writing, each grammar and style recommendation was considered individually.

To mitigate risks, control was maintained over all final decisions. All code implementations were thoroughly tested, synthetic data was verified to represent intended scenarios accurately, and all writing suggestions were reviewed before adoption. The analysis, conclusions, and technical implementation represent the authors' own work, with AI serving as a supportive tool rather than a primary source of content.

References

Banon (2010) “Elk Stack.” Amsterdam, NL. 2012. *Elastic Stack*, ver. 9.0.4. Available at: https://www.elastic.co/guide/index.html.

Baum, Das, and Swan (2003) “Splunk.” San Jose, CA, USA. 2003. *Splunk Machine Learning Toolkit*, ver. 9.4. Available at: https://help.splunk.com/en.

Buczak and Guven (2015) “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” *IEEE Communications Surveys & Tutorials*, 18(2), pp. 1153–1176. Available at: https://doi.org/10.1109/COMST.2015.2494502.

Director, Operational Test and Evaluation (2024) *Cyber Assessment Program*. Washington, DC, USA. Available at: https://www.dote.osd.mil/Portals/97/pub/reports/FY2024/dotemanaged/2024cap.pdf?ver=CYXAHEh1-0JMR-9Jde33wA%3D%3D.

Kennedy (2010) “Social Engineering Toolkit.” Fairlawn, OH, USA. 2010. *Social Engineering Toolkit*, ver. 8.0.3. Available at: https://github.com/trustedsec/social-engineer-toolkit.

Koopmann (2010) “Graylog.” Houston, TX, USA. 2009. *Graylog*, ver. 6.6. Available at: https://graylog.org.

Lyon (1997) “Network Mapper.” San Jose, CA, USA. 1997. *Network Mapper*, ver. 7.97. Available at: https://svn.nmap.org/!svn/bc/3320/nmap/docs/nmap_manpage.html≠.

Moore (2003) “Metasploit.” Boston, MA, USA. 2003. *Metasploit*, ver. 4.22.8. Available at: https://www.metasploit.com.

National Institute for Standards and Technology (2025) *Security and Privacy Controls for Information Systems and Organizations*. Rep. 800-53, rev. 5. Gaithersburg, MD, USA. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-53r5.pdf.

Novikov (2017) *Log Analysis*, *Wallarm*. Available at: https://www.wallarm.com/what/log-analysis.

Raval, V. and Verma, S. (2017) “Challenges of Security Log Management,” *ISACA Journal*, 6. Available at: https://www.isaca.org/resources/isaca-journal/issues/2017/volume-6/the-practical-aspect-challenges-of-security-log-management.

Schroeder and Ross (2015) “PowerShell Empire.” Las Vegas, NV, USA. 2010. *Powershell Empire*, ver. 6.1.3. Available at: https://github.com/BC-SECURITY/Empire.

Tariq *et al.* (2025) “Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities,” *ACM Computing Surveys*, 57(9), pp. 1–38. Available at: https://doi.org/10.1145/3723158.

A Key Rotation Management System: Design and Implementation for Improved Data Security

Malibongwe Ntalali[1], Noluntu Mpekoa[1] and Sheethal Tom[2]
[1]University of Johannesburg, South Africa
[2]The Independent Institute of Education, Varsity College, Cape Town, South Africa

223081688@student.uj.ac.za
noluntum@uj.ac.za
stom@varsitycollege.co.za

Abstract: Organizations in today's rapidly changing digital world use various channels, such as secure APIs and encrypted communications, to enhance collaborations and information sharing. While these systems assist in enhancing productivity, innovation and collaboration, they can also introduce major security risks for protecting sensitive information. The core of data protection depends on cryptographic key management, with key rotation playing a vital role, yet often neglected. Key rotation, a well-established cryptographic practice, is essential for minimizing data exposure, mitigating key compromise risks and ensuring regulatory compliance. However, inconsistent implementation of key rotation policies across organizations often results in varied security practices. The proposed Key Rotation Management System functions as a solution that automates and standardizes all stages of the key lifecycle process. The system implements a three-tier Model-View-Controller framework which combines different functional components that include user authentication together with role-based control, automated key generation, secure storage and distribution, periodic rotation, dashboard visualization and proactive alert systems. The proposed system resolves fundamental problems which arise from human involvement in administration and irregular scheduling and insufficient user understanding and insecure distribution methods. The solution provides real-time key ageing visibility through its dashboard interface, while its scheduling and reminder features assist in automated rotation interval execution. It includes a centralized service request-response module together with automated email notification to help organizations maintain effective communication and monitor compliance standards. The evaluation results show that the proposed system enhances user responsiveness and decreases expired key occurrences while decreasing manual work and assisting organizations in meeting industry international standards. The system design with role-based access controls provides both security measures and system accountability features. The proposed Key Rotation Management System combines automation, visualization and security to offer a scalable solution that can strengthen the cryptographic strength of digital infrastructures.

Keywords: Key rotation, Cybersecurity automation, Cryptographic key management, Secure data sharing, Key lifecycle management

1. Introduction

Organizations use technologies such as emails, video conferencing and intranet for collaboration (Hurst, 2020). With the increase in collaboration and information sharing, the security of sensitive information has become paramount. Organizations have taken additional precautions to store sensitive information such as passwords, tokens, and private keys in centralized key storage locations, procuring them only during runtime (Ahanger et al, 2024). With the storage and usage of this sensitive information, one critical aspect of maintaining this security is the effective management of cryptographic keys (Ahmad, Mehfuz and Beg, 2023). Key management, structured as a lifecycle includes the secure generation, storage, distribution, use, and destruction of keys (Faisal et al, 2020). The Open Web Application Security Project (OWASP) recommends best practices such as key generation within cryptographic modules, distributing them securely and never storing them in plain text (Sakthivel et al, 2020).

Key rotation, the process of periodically changing encryption keys, is essential to mitigate the risks associated with key compromise and ensure continued data protection. Key rotation is mandated by regulation in some contexts, such as the Payment Card Industry Data Security Standard (PCI DSS) that dictates how credit card data must be secured [PCI16] (Seaman, 2020). Despite its importance, many organizations face challenges with implementing efficient key rotation strategies due to the complexity and resource requirements involved. The most challenging step in the key management life cycle is key rotation, as it requires change management, stakeholder adoption and management, policy enforcement and coordination across different platforms. The challenges faced by the cloud users with key management to securely interact with a range of services and store data created or processed by those services are due to differences in ownership and control of the underlying infrastructures on which the protected resources and the key management system (KMS) are located (Ahmad et al, 2025).

This study aims to address these challenges by proposing a comprehensive Key Rotation Management System (KRMS) that automates and streamlines the key rotation process, thereby enhancing security and reducing operational overhead. The system will serve as a centralized hub for control and communication, facilitating the scheduling of notifications and reminder alerts. Additionally, it will enable the monitoring of cryptographic key aging and provide a comprehensive dashboard for oversight and management.

The remainder of the paper reviews related work (Section 2), introduces the KRMS model and architecture (Section 3), details its implementation (Section 4), presents results and discussion (Section 5) and concludes with key contributions (Section 6).

2. Literature Review

This section focuses on Key Rotation Management System (KRMS) relevant studies, starting with broader aspects of information security before moving to more specific aspects that are related to the proposed system.

2.1 Background on Information Security

Information security protects the confidentiality, integrity, and availability (CIA) of information assets (Mpekoa, 2024). Confidentiality is achieved through encryption, which prevents unauthorized access (Zhang et al, 2018). The protection of data integrity involves protecting it against unauthorized damage, disruption or alteration to maintain its accurate and trusted state. Availability ensures that authorized users, including people and systems maintain access to information during their required times while receiving it in a usable form (Yee and Zolkipli, 2021). The three fundamental principles work together to create a resilient and robust network security framework.

Human beings are the most vulnerable point of information security, as many breaches stem from human errors or carelessness (Hughes-Lartey et al,2021). Hence, Security Education, Training, and Awareness (SETA) is imperative when implementing a security system in any organization. SETA raises the awareness of protecting organizational information, provides employees with the skills and knowledge to work securely and creates a shared organizational culture of security (Bethel, 2020).

2.2 Cryptography and Key Management

Cryptography ensures confidentiality, integrity, authentication, and availability through the use of encryption algorithms. Encryption is a mathematical process used to transform text into a coded format, making it unreadable without the corresponding decryption method. Only those who possess knowledge of the algorithm and all the necessary key material utilized in the process can decode the encrypted text. Encryption algorithms can be categorized as either symmetric or asymmetric algorithms. In a symmetric algorithm, the sender and receiver use a shared secret key for encryption and decryption (Yassein et al, 2017). While this method ensures message privacy, it poses security risks during key distribution. Thus, it is crucial to find a secure way to distribute and protect the key. Asymmetric algorithms use a pair of mathematically related keys instead of a single key. The sender encrypts the message with a public key, while the receiver decrypts it using a private key, known only to them. Public keys can be shared widely, while private keys remain confidential (Hinkes, 2018).

The effectiveness of cryptography relies on strong key management because quantum-resistant primitives are currently emerging. Strong algorithms become weak when poorly managed keys are handled by poor practices (Barker and Barker, 2018). A key management system handles all aspects of cryptographic key operations from secure generation to storage, distribution, usage and eventually key retirement. According to NIST's key management guidelines (SP 800-57 revised 2024), organizations must establish cryptoperiods and perform regular key rotation to reduce the exposure time in case of key compromise (Vlajic and Cianfarani, 2023). Studies (Torres, 2022; Ahmad, Mehfuz and Beg, 2023) emphasize the necessity of automated key lifecycle management in hybrid and cloud-based systems.

2.3 Challenges in Key Rotation

The requirement for key rotation in PCI DSS compliance standards creates major obstacles during implementation. Manual processes in key rotation are error-prone and tend to result in non-compliance and missed rotation deadlines (Kuzminykh, Yevdokymenko and Ageyev, 2020). Research indicates that cloud and hybrid infrastructure key rotation face challenges because of ownership fragmentation and interoperability issues. The "key sprawl" phenomenon becomes more prevalent according to research because it rates difficulties for service orchestration while increasing service disruption risks (Nakamura et al, 2025). The

cryptoperiods should be adaptive, but the practicality of this remains in question due to organizational resource constraints and integration difficulties (Kim et al, 2016).

2.4 Need for Key Rotation Management System

Key management encompasses a comprehensive set of techniques aimed at the initialization, registration, updating, and recovery of cryptographic keys. These techniques are crucial for maintaining the confidentiality, integrity, and authentication of communications among authorized entities (Gautam and Kumar, 2021). However, current commercial solutions and academic literature often overlook the critical aspect of key rotation within the broader framework of key management. The implementation of key rotation policies tends to vary significantly across different organizations, leading to inconsistent security practices that can create vulnerabilities. This disparity highlights a pressing need for standardized approaches to key rotation that can enhance overall security (Joshi, Crowther and Robinson, 2024). Additionally, many existing solutions lack straightforward automation features, particularly those that provide proactive notifications when keys are nearing expiration or require rotation. Furthermore, the literature indicates a significant shortage of expertise and robust systems capable of facilitating an effective lifecycle of key management and compliance monitoring. This gap not only hampers organizations' ability to manage keys efficiently but also increases the risk of security breaches due to outdated or improperly managed cryptographic keys (Radanliev, 2023). Addressing these challenges is essential for organizations seeking to strengthen their security posture in an increasingly complex threat landscape.

3. System Architecture and Model Implementation

The proposed Key Rotation Management System (KRMS) is a comprehensive framework designed to efficiently handle the lifecycle of cryptographic keys. The system encompasses several critical modules, each of which plays a vital role in ensuring robust key management practices. The proposed KRMS implements a Model-View-Controller (MVC) architecture to ensure separation of concerns, maintainability, modularity and scalability. It makes a codebase's concern separation easier by breaking it up into three separate layers:

- the *Model*, which takes care of data management and business logic;
- the *View*, which handles user input and output presentation; and
- the *Controller*, which acts as a mediator between the Model and the View, managing interactions and preserving the application’s state.

This structure makes code easier to organize and more maintainable (Nahhas, 2021). The three-layer architecture is illustrated in Figure 1 below.

The Presentation Layer provides a user interface through a Flask-based web application. The responsibilities of this layer include rendering web pages based on user requests, collecting and validating user inputs before passing them to the business logic layer, and displaying results or errors based on responses from the business logic layer. The Business Logic Layer contains the fundamental KRMS operations, which generate keys automatically and schedule rotations and enforce security protocols. The system uses this layer to link user requests with system operations and maintain compliance with cryptographic standards. The data access layer enables safe database interactions with the KRMS system. The data access layer communicates with the business layer to obtain database instructions, which it then uses to communicate with a MySQL database via a Flask-based Python application using a MySQL connector.

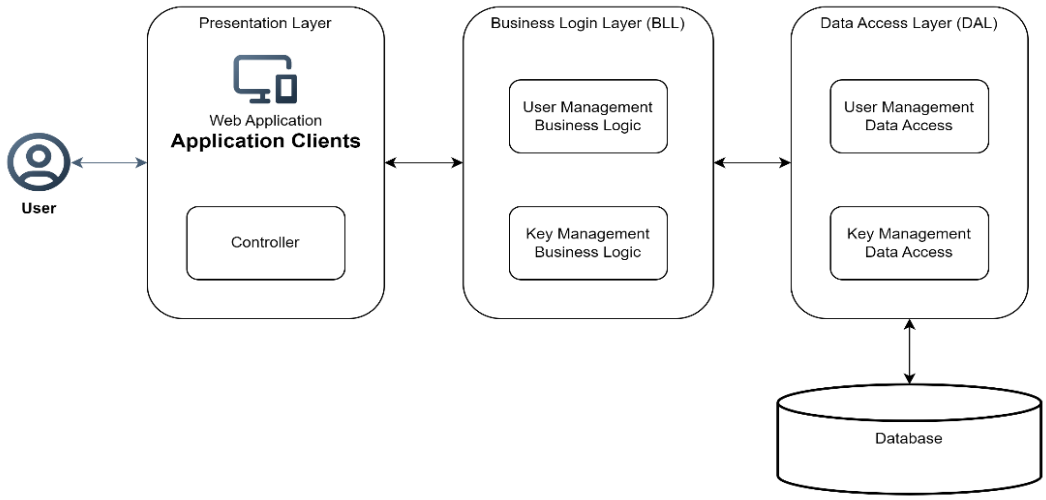


Figure 1: Key Rotation Management System Architecture Diagram

The layered design of KRMS allows organisations to monitor cryptographic keys from a central point while providing features like automation and proactive notification capabilities.

4. Proposed Solution - Key Components

The KRMS provides automation alongside visualization tools and notification features through an adaptable centralised platform. The KRMS implements practical and scalable lifecycle management through user-friendly interfaces which connect security features to usability. The key modules within the KRMS are outlined and described in detail as follows:

4.1 User Management Module

The User Management Module is a critical component of the KRMS that plays a central role in overseeing the creation, management, and authentication of user accounts within the system. Its primary purpose is to ensure that only authorized individuals can access sensitive operations related to key management. This module handles authentication (verifying user identity), authorization (enforcing role-based access controls) and user management (creation, updating, and deletion of user accounts). Key features of the module include user registration with secure handling of passwords (hashing and salting), login-based user authentication and Role-Based Access Control (RBAC) with defined user roles (e.g., Admin, User). User profile management is another crucial feature, enabling users to view and update their profiles, including personal information and security settings, and providing functionalities for password changes and account recovery. Finally, the module supports account deactivation and deletion, ensuring that user accounts can be deactivated or removed when no longer needed, in compliance with data retention policies.

4.2 Key Management Module

The Key Management Module ensures key generation, distribution, storage, rotation, and retirement. Key generation is handled by utilizing the Python secrets library, which is designed to generate cryptographically secure pseudo-random numbers. This library ensures that the keys used for managing secrets, such as authentication tokens or API tokens, meet the necessary length and complexity standards. By configuring the token key length appropriately, the generated keys adhere to the required security measures. In terms of key distribution, it is essential to ensure that keys are decrypted in a secure environment to prevent unauthorized access. The decryption process reverses the encryption steps: it starts by decoding the Base64-encoded encrypted key into bytes, verifies the secret key, and if correct, extracts the initialization vector (IV) and encrypted data. An AES cipher is then used to decrypt the data, removing any added padding to retrieve the original key value, which is decoded from bytes to a UTF-8 string. The KRMS enables secure key availability through its rotation system which protects keys from interception. The storage process uses AES-CBC encryption to safeguard the keys, which are then combined with IVs prior to Base64 encoding for storage and transmission with error handling mechanisms. The key rotation process follows established policies through automated systems, which minimizes human errors. Through decommissioning, deletion or archiving techniques, the retirement process safely eliminates obsolete keys; however, in the event of a breach, manual intervention is necessary.

4.3 Service Request - Response Module

The Service Request/Response Module within the KRMS serves as an essential communication channel, facilitating collaboration between clients and administrators. This module is designed to streamline and manage the exchange of information and requests related to key management and system operations. This module enables clients to submit detailed requests to administrators. The module produces both system updates and automatic email notifications to keep administrators immediately aware of each request. The KRMS allows administrators to review requests while clients receive email updates that also get logged in the module for future reference. This module gives administrators a controlled system to handle client requests for support. The module maintains a clear record of all responses that administrators make. The system automatically generates an email notification to clients after submission, which enables prompt and transparent documentation for efficient issue resolution.

4.4 Reminder Capability Module

The KRMS reminder function enables administrators to schedule key updates through rotation schedule-based reminders, which help maintain timely key updates. The system adjusts schedules based on defined parameters, which include timing, key groups and notification content that can be modified as needed. Additionally, reminders can be removed from the system once they have served their purpose. This includes automatic removal after a reminder is sent, as well as manual deletion of outdated or irrelevant notifications. This automated cleanup helps maintain an organized and efficient reminder system.

4.5 Dashboard and Monitoring Module

The Dashboard and Monitoring Module in the KRMS is a vital feature that provides administrators and clients with a comprehensive overview of key statuses and ageing. This module is designed to facilitate efficient key management by visually categorizing keys based on their lifecycle stages, allowing for timely key rotations and proactive maintenance.

Dashboard Overview

The dashboard offers a summary view of the key statuses, using a color-coded system to differentiate between various stages of key validity:

- **Valid Keys (Green):** Keys that are currently active and within their designated validity period are displayed in green. The green indicator helps administrators quickly identify which keys are currently operational and in good standing.
- **Close to Expiry (Amber):** Keys that are approaching their expiration date are highlighted in orange or amber. The amber alert allows administrators to plan ahead and initiate key rotation processes before these keys expire, thereby avoiding potential disruptions or security issues.
- **Expired Keys (Red):** Keys that have surpassed their validity period and are no longer usable are marked in red. The red status ensures that expired keys are not used inadvertently and prompts administrators to take necessary actions, such as removing or replacing the expired keys.

Monitoring and Alerts

The monitoring aspect of the module extends beyond visual status indicators to include dynamic tracking and alerting functionalities:

- **Real-Time Monitoring**: The module continuously updates key statuses in real-time, providing administrators with the most current information. This real-time tracking ensures that administrators are always aware of the current condition of each key.
- **Alerts and Notifications**: Automated alerts are configured to notify administrators when keys are approaching expiration or have expired. These alerts can be customized based on specific thresholds and criteria, ensuring that administrators receive timely reminders to take appropriate actions.

4.6 Scheduler Module

The Scheduler Module is responsible for automated key rotation through scheduled alerts which are uniformly applied across systems and services, thereby reducing the risk of key compromise. The Scheduler Module encompasses key functions such as defining the frequency of key rotations—whether daily, weekly, or monthly—and allowing for customization to set different rotation schedules based on the type of key or service and its criticality. Additionally, it includes features for setting reminders and alerts, which automatically notify

stakeholders of upcoming key rotations and provide configurable reminders to facilitate timely preparation and execution of key rotation tasks.

5. Results and Discussion

Usability testing was not conducted on the proposed solution; instead, the focus was exclusively on functional testing to evaluate its performance and reliability. Functional testing involves assessing specific features and functionalities to ensure they operate as intended under various conditions. The results of this testing are detailed in the sections below.

5.1 System Functionality

The functional testing proved that the KRMS executed cryptographic key rotation based on established schedules. Figure 2 below displays the Key Generation and setup screen. The user enters the service name, selects the key administrator, key type, key value, rotation frequency and lastly, schedule type.

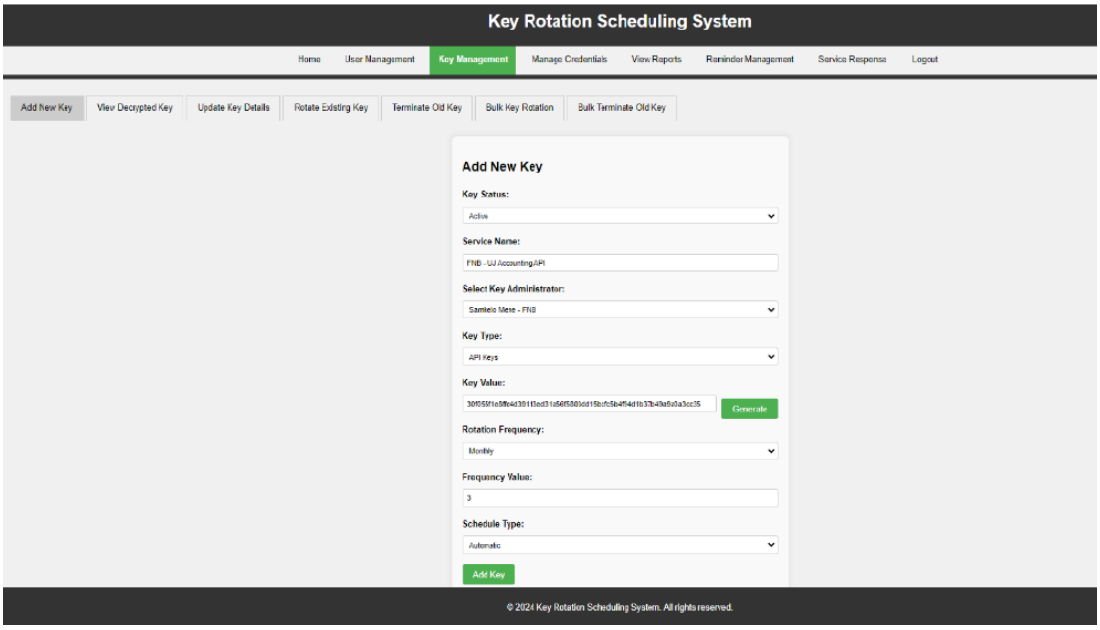


Figure 2: Key Generation and Setup

Figure 3 below provides a real-time overview of key statuses ('Valid', 'Close to Expiry', 'Expired'). But over and above this available facility, the system automates email notifications and directly informs users about key statuses, prompting timely actions and adherence to security protocols.

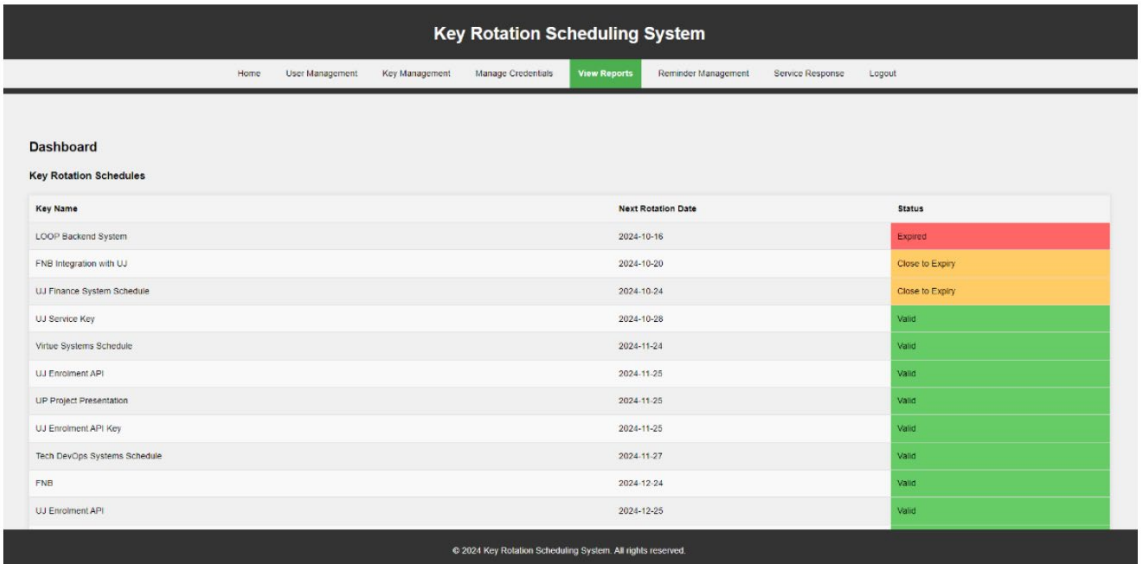


Figure 3: Key Rotation Scheduling

These outputs show that the KRMs provide real-time visibility as well as proactive alerts and verifiable evidence of compliance. The system maintained consistent logs of all lifecycle events, including key generation, storage, rotation and termination, which created a system of accountability while minimizing audit ambiguities.

5.2 System Interaction

The keys are stored in an encrypted form as illustrated in Figure 4 below. It is essential to ensure that keys are encrypted and decrypted in a secure environment to prevent unauthorized access.

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Content:

id	expiration_date	System_Service_Name	key_type	new_key	old_key	key_status	rotated_datetime	schedule_status
39	2024-10-16	LOOP Backend System	RSA	0TUVzFulAer26s1+XCvqCULrln0ripv6wtaKp…	NULL	1	2024-10-16 09:41:28	manual
40	2024-11-24	Virtue Systems Schedule	Encryption Keys	FDGXhXiC8EqhpaAbx70eeZSyvXZ6GEqjMPo…	NULL	1	2024-09-24 14:08:53	automatic
42	2024-11-27	Tech DevOps Systems Schedule	API Keys	REoLcwc5MgjsEoxU/HQHOPd1KiJPs7x4JshQ…	NULL	1	2024-08-28 15:48:33	automatic
44	2025-09-24	Virtue Systems OPT	CORE	xu6VpZ/NhMwypJwoAmhADCy+b8bf+k+Me…	NULL	1	2024-09-24 17:34:08	manual
45	2025-07-28	InfoSec Team Service	CORE	Ly4lvZLvWh1VZS3yX3xp8/CUiwR5UCuZqlDV…	NULL	1	2024-08-28 15:48:33	manual
47	2024-10-24	UJ Finance System Schedule	Asymmetric Keys	u2D/9y4GTiq7HavfUMhwGQDmaKJ8nNUoz…	NULL	1	2024-09-24 14:09:01	automatic
48	2024-10-28	UJ Service Key	API Access Key	wRPNaz5kHYCFTWTj84I5OEj1oh9v4WS+/y…	NULL	1	2024-08-28 19:45:46	manual
49	2025-10-15	UP Loyalty Service Schedule	Symmetric Keys	TYboKm5w/tARlZDFuUIVqOu2tvldv+qEZBn…	NULL	1	2024-10-15 20:17:34	automatic
50	2025-08-22	Virtue Systems OPT	CORE	fd+UnrWkbjuPbe+Lt1nMDcU0KR6QN7N/NI…	NULL	1	2024-08-28 15:48:33	manual
51	2024-11-25	UJ Enrolment API	ACCESS	483BM6060TXKJSDILt8Nxarje603GSxVHNP8…	NULL	1	2024-09-25 08:46:09	manual
52	2024-12-24	FNB	API Access Key	qpNk5bukTPtumNSchQGO0CotAX3Wd+OjL9…	NULL	1	2024-09-24 17:32:59	manual
53	2024-12-25	UJ Enrolment API	API Access Key	5jxvHhnOs9BGzm6sRIYjN7UsVQDG8Au+fYD…	NULL	1	2024-09-25 08:40:32	manual
54	2024-11-25	UP Project Presentation	API Access Key	LNTDRsnmY+Z+7yb9il0cRvRUIK58uVjC6JDJ…	NULL	1	2024-09-25 09:42:44	manual
56	2024-11-25	UJ Enrolment API Key	API Access Key	yYqvUMsnZwH5m/CAYyeqy8KF62jgN9CFbi2…	NULL	1	2024-09-25 11:18:29	manual
57	2025-10-20	UJ Enrolment API with Home Affairs Sch…	API Keys	xfbCAx6Kokx34a5qEOyjrfvF74smJd4FhSxR…	qBjHlYA9rcnZ0POb+LoVN7uSU8TDohbi…	1	2024-10-20 17:50:16	automatic
60	2025-01-16	UJ Enrolment API with Home Affairs	API Keys	FDPxED3YRiYLXqjgZQV7+Uc2s3GuU9KVrSp…	NULL	1	2024-10-16 09:40:22	automatic
62	2024-10-20	FNB Integration with UJ	Session Keys	NBKj0BMbLESVCbUulYIwzRb/H3xdsAMJVN0s…	8f9x4de6gHWKThEO0AhmgiVmeCBWx…	1	2024-10-20 17:00:25	automatic
NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

Figure 4: Cryptographic Key Storage

The system achieved real-time capabilities alongside lifecycle visibility (key generation, storage, distribution, rotation, and termination) by using multiple channels that supported immediate reaction and long-term oversight. The proposed KRMS provides a centralised platform for all system communications, which eliminates the requirement for administrators to track multiple information streams. The stream implements role-based views that grant administrators and super administrators access to appropriate tools and information based on their job responsibilities.

5.3 System Validation and Reliability

The functionality testing process confirmed that the KRMS executed all its fundamental lifecycle operations as specified in the design. The distribution process proved to be streamlined, allowing for safe and efficient key sharing with authorized parties. Key rotation was automated, reducing manual intervention and minimizing the risk of human error. Finally, the key termination process guarantees that expired or compromised keys are promptly and securely deactivated, enhancing overall data protection. Collectively, these capabilities contribute to the robust security framework provided by the KRMS, ensuring that all aspects of key management are effectively handled.

5.4 Critique and Analysis

The automation of key rotation greatly enhances efficiency, reducing the risk of human error and ensuring keys are rotated at appropriate intervals without manual intervention. This consistency is crucial for adhering to key rotation security policies. The proactive notification system, such as email reminders, helps inform users of upcoming expirations and necessary actions, fostering a culture of security awareness and compliance. Audit trails also help meet regulatory requirements, providing evidence that key management policies are being followed. Error handling further improves the system's reliability by managing failures gracefully, reducing the likelihood of complete system shutdowns, and ensuring smoother operations. The results of this study proved technical feasibility, yet several areas require improvement.

The lack of integration with calendars and the selection of business days creates inefficiencies in scheduling. Relying on hard-coded start dates for tasks makes adapting the system to different environments challenging, necessitating frequent updates or maintenance. Limited scalability is another concern; performance might degrade under high volumes of reminders or key rotations, indicating potential bottlenecks. Ambiguities in handling time zones complicate scheduling tasks across different geographical locations, especially in global organizations. Lastly, the reliance on a single instance of the scheduler presents a vulnerability—if that instance fails, all scheduled tasks may not execute.

To address these issues, future research could explore more advanced scheduling algorithms that adapt to load and resource availability, improving efficiency and responsiveness. Understanding user behaviour through analytics could provide insights into enhancing the user experience and identifying potential security risks. Integrating the KRMS with other security frameworks, such as Security Information and Event Management

(SIEM) systems, could strengthen the overall security posture and incident response capabilities. SIEM will improve the KRMS because it can collect, store, aggregate, and correlate events and information generated by a managed system infrastructure. Developing smarter notification systems that adapt based on user preferences and behaviours may increase engagement and compliance rates. Additionally, incorporating more detailed reporting and analytics capabilities would allow for better insights into key management activities and trends.

6. Conclusion

Organizations must maintain effective key management systems because the changing information security environment demands protection of sensitive data during collaborative technology adoption. The proposed KRMS solves key rotation problems by implementing a systematic, automated system for cryptographic key management throughout its entire existence. The KRMS uses automated integration with existing IT systems to simplify key rotation processes while reducing human involvement and minimizing errors. The Model-View-Controller pattern-based layered architecture provides scalability alongside maintainability and clear separation of concerns. Key components of the KRMS, including the User Management Module, Key Management Module, Service Request/Response Module, Reminder Capability Module, Dashboard and Monitoring Module, and Scheduler Module, each play a crucial role in enhancing the security and efficiency of key management practices. These components collectively support robust key lifecycle management by facilitating secure key generation, distribution, storage, rotation, and termination. Additionally, the system's reminder and scheduling features help maintain compliance with regulatory standards and organizational policies, ensuring timely key updates and minimizing security risks.

The study found that the KRMS effectively addresses the initial challenges identified, including the necessity for manual administrator intervention for key updates, a lack of awareness about key rotation processes, inadequate monitoring of key age, insecure key distribution methods, and insufficient communication channels. The implementation of a single point of communication, automated user notifications, and automated key rotation has significantly improved user engagement, efficiency, security, user satisfaction, and awareness of key management processes. The KRMS demonstrated high success rates in minimizing human error and enhancing operational efficiency through automated key rotations. While initial qualitative feedback is positive, further research with a broader user base is necessary to fully validate these findings. KRMS demonstrates through its automated key lifecycle management that organisations can achieve better security. The KRMS offers an extensible automated key rotation system which enhances information security practices and enables organisations to handle rising collaborations and technology complexity.

Acknowledgements

Not Applicable

Ethics Declaration: Ethical clearance was not required for this study as it did not involve human participants.

AI declaration: Grammarly was utilized for editing and grammar refinement.

References

Ahanger, A.S., Masoodi, F.S., Khanam, A. and Ashraf, W. (2024) Managing and securing information storage in the Internet of Things, in Internet of Things Vulnerabilities and Recovery Strategies, Auerbach Publications, pp. 102–151.

Ahmad, S., Mehfuz, S. and Beg, J. (2023) "Hybrid cryptographic approach to enhance the mode of key management system in cloud environment", The Journal of Supercomputing, Vol. 79, No. 7, pp. 7377–7413.

Ahmad, S., Nazim, M., Arif, M., Ahmad, J., Mehfuz, S. and Ansari, M.A. (2025) "Protecting data in the cloud: A systematic literature review of key management", Concurrency and Computation: Practice and Experience, Vol. 37, Nos. 21–22, p. e70223.

Barker, E. and Barker, W. (2018) Recommendation for Key Management, Part 2: Best Practices for Key Management Organization (Draft), NIST Special Publication (SP) 800-57 Part 2 Rev. 1, National Institute of Standards and Technology, Gaithersburg.

Bethel, K.L. (2020) An Evaluation of Organizational Culture: Its Influence on Security Culture: A Case Study, Doctoral dissertation, Northcentral University.

Faisal, M., Ali, I., Khan, M.S., Kim, J. and Kim, S.M. (2020) "Cyber security and key management issues for internet of things: Techniques, requirements, and challenges", Complexity, Vol. 2020, No. 1, p. 6619498.

Gautam, A.K. and Kumar, R. (2021) "A comprehensive study on key management, authentication and trust management techniques in wireless sensor networks", SN Applied Sciences, Vol. 3, No. 1, p. 50.

Hinkes, A.M. (2018) "Throw away the key, or the key holder? Coercive contempt for lost or forgotten cryptocurrency private keys, or obstinate holders", Northwestern Journal of Technology & Intellectual Property, Vol. 16, pp. 225.

Hughes-Lartey, K., Li, M., Botchey, F.E. and Qin, Z. (2021) "Human factor, a critical weak point in the information security of an organization's Internet of Things", Heliyon, Vol. 7, No. 3.

Hurst, E.J. (2020) "Web conferencing and collaboration tools and trends", Journal of Hospital Librarianship, Vol. 20, No. 3, pp. 266–279.

Joshi, S., Crowther, K. and Robinson, J. (2024) "Tradeoffs in key rotation strategies for industrial internet of things devices and firmware", Applied Sciences, Vol. 14, No. 21, p. 9942.

Kim, H., Wasicek, A., Mehne, B. and Lee, E.A. (2016) "A secure network architecture for the internet of things based on local authorization entities", in 2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud), IEEE, pp. 114–122.

Kuzminykh, I., Yevdokymenko, M. and Ageyev, D. (2020) "Analysis of encryption key management systems: strengths, weaknesses, opportunities, threats", in 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S&T), IEEE, pp. 515–520.

Mpekoa, N. (2024) "An analysis of cybersecurity architectures", in International Conference on Cyber Warfare and Security, Academic Conferences International Limited, March, pp. 200–207.

Nahhas, S. (2021) "MVC architecture from maintenance quality attributes perspective", International Journal of Computer Science and Security (IJCSS), Vol. 15, No. 5, pp. 169–181.

Nakamura, Y., Fukuda, T., Yang, X. and Takefuji, Y. (2025) "SSHAA: A Python package index for visualizing features of SSH attacks with text mining in classification", Information Security Journal: A Global Perspective, Vol. 34, No. 1, pp. 50–62.

Radanliev, P. (2023) Cyber-attacks on Public Key Cryptography.

Sakthivel, M., Sivanantham, S., Bharathiraja, N., Krishna, N.B., Kamalraj, R. and Kumar, V.S. (2024) "Ensuring web application security: An OWASP driven development methodology", in 2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS), IEEE, April, Vol. 1, pp. 1–7.

Seaman, J. (2020) PCI DSS: An Integrated Data Security Standard Guide, Apress.

Torres, A.G. (2022) "Encryption key lifecycle management and best practices for maintaining trusted e-commerce services in the cloud", Journal of Artificial Intelligence and Machine Learning in Cloud Computing Systems, Vol. 6, No. 11, pp. 1–8.

Vlajic, N. and Cianfarani, G. (2023) "Risk-based methodology for optimal cryptoperiod calculation in ICSs under data siphoning attack", in Proceedings of the 2024 Workshop on Re-design Industrial Control Systems with Security, November, pp. 87–95.

Yassein, M.B., Aljawarneh, S., Qawasmeh, E., Mardini, W. and Khamayseh, Y. (2017) "Comprehensive study of symmetric key and asymmetric key encryption algorithms", in 2017 International Conference on Engineering and Technology (ICET), IEEE, August, pp. 1–7.

Yee, C.K. and Zolkipli, M.F. (2021) "Review on confidentiality, integrity and availability in information security", Journal of ICT in Education, Vol. 8, No. 2, pp. 34–42.

Zhang, Q., Jia, S., Chang, B. and Chen, B. (2018) "Ensuring data confidentiality via plausibly deniable encryption and secure deletion – a survey", Cybersecurity, Vol. 1, No. 1, p. 1.

Explainable AI in Insider Financial Fraud Detection Models: A Review of Transparency and Trust

Hillary Kwame Ofori[1], William Leslie Brown-Acquaye[1], Forgor Lempogo[1], Kwame Bell-Dzide[1] and Israel Edem Agbehadji[1,2]

[1]Ghana Communication Technology University, Accra North, Ghana
[2]Durban University of Technology, Durban, South Africa

hofori@gctu.edu.gh
wbrown@gctu.edu.gh
lforgor@gctu.edu.gh
kbell-dzide@gctu.edu.gh
israeldel2006@gmail.com

Abstract: Financial and insider fraud increasingly intersect with broader cybercrime ecosystems, creating attack vectors that undermine national cyber resilience and the integrity of digital financial infrastructures. As organizations turn to machine learning (ML) and deep learning (DL) models for automated fraud and insider-threat detection, the opacity of these systems presents strategic risks for cyber defense: unexplainable alerts weaken analyst trust, complicate incident response, and challenge regulatory and forensic accountability. This study presents a systematic review of 107 empirically validated works (2015–2025) examining how Explainable Artificial Intelligence (XAI) techniques enhance transparency, trustworthiness, and operational readiness in AI-driven fraud detection systems. Using a mixed bibliometric–thematic methodology, the review maps the evolution of ML/DL architectures, XAI adoption patterns, evaluation practices, and dataset limitations within security-critical environments. The findings highlight a sector-wide dependence on post-hoc feature attribution and reveal emerging shifts toward intrinsic interpretability through attention mechanisms and hybrid temporal models. Despite progress, gaps persist: limited use of sequential behavioral models, narrow evaluation metrics, and overreliance on structured datasets weaken real-world resilience against adaptive adversaries. To address these challenges, the paper proposes a Three-Pillar Framework: Algorithmic Transparency, Evaluation Accountability, and Data Traceability that positions explainability as a foundational architectural property for cyber defense systems. By aligning model interpretability with security operations, regulatory requirements, and analyst cognition, the framework strengthens organizational readiness against insider threats, financial fraud, and AI-targeted adversarial manipulation, key considerations in modern cyber warfare and security operations.

Keywords: Explainable AI, Transparent model, Model interpretability, Financial fraud detection, Insider fraud detection

1. Introduction

Financial fraud, insider threats, and transactional manipulation have become key attack methods in modern cyber warfare and cybercrime networks. As national economies rely more on digital financial infrastructure, adversaries exploit large-scale transactional processes to disrupt operations, steal funds, and undermine public confidence (Tian, 2025; Zhu and Chen, 2024). These developments elevate fraud detection from a financial risk issue to a matter of cybersecurity and national security, as emphasized by regulatory agencies and global security groups (Adetumi et al., 2024; Zhu et al., 2025). In this environment, organizations are increasingly deploying machine learning (ML) and deep learning (DL) models to detect complex, adaptable threats that traditional rule-based systems often miss (Aljunaid et al., 2025; Zhang et al., 2025).

However, the widespread use of opaque ML/DL models introduces new vulnerabilities in cyber defense. Black-box fraud detection systems hide decision logic precisely when analysts and incident responders need actionable intelligence. This tension is well-understood in socio-technical theory, which stresses that technological outputs must remain interpretable to the human defenders responsible for operational responses and risk decisions (Al-hchaimi et al., 2024; Qin et al., 2022; Roshan and Zafar, 2021). In cyber defense environments, where forensic traceability, adversary attribution, and rapid triage are critical, unexplainable model outputs damage trust, slow response coordination, and threaten regulatory compliance.

Explainable Artificial Intelligence (XAI) has become a key tool in cybersecurity, enhancing both transparency and mission preparedness. Foundational work by Gunning and Aha (2019) and the development of Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) methods (Ribeiro, Singh, and Guestrin, 2016) highlight the importance of explainability for making AI-driven predictions auditable, verifiable, and meaningful in operations. Regulations like the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence (EUAI) Act (The European Parliament, 2016, 2024) reinforce these needs through legal requirements for transparency and accountability, especially crucial for

national financial systems targeted by cyber threats. Recent studies also show that transparent Artificial Intelligence (AI) improves security by increasing analyst confidence, reducing false positives, and strengthening human-AI teamwork in high-stakes decision-making (Abbas, Hilal, and Jabbar, 2025; Praveenraj et al., 2023; Vivek et al., 2025).

In parallel, emerging explainability-driven architectures such as eXplainable Security Policy-Induced Data Engineering (X-SPIDE) for smart contract fraud (Pennella, Pinelli, and Galletta, 2025) and hybrid Bidirectional Long Short-Term Memory – Variational Autoencoder (BiLSTM-VAE) models for behavioral anomaly detection (Ofori et al., 2025) reflect a shift toward security-oriented, intrinsically interpretable systems. From the perspective of cognitive fit theory, these approaches better support the mental models of cybersecurity analysts; enabling them to identify anomalous sequences, validate alerts, and anticipate adversarial behavior.

Despite these advancements, significant gaps still exist in cybersecurity research. Many studies on fraud and insider-threat detection continue to focus on predictive accuracy rather than interpretability, making analysts rely on opaque reasoning processes. There is limited evidence on which XAI techniques offer the most stable, reliable, and domain-appropriate explanations in high-stakes cyber defense settings (Vivek et al., 2025). Furthermore, the prevalence of post-hoc explainers like SHAP and LIME often raises concerns about explanation fidelity and their vulnerability to adversarial attacks (Agomuo et al., 2025). The lack of standardized methods and fragmented evaluation practices makes it harder for security operations centers (SOCs), auditors, and regulators to adopt transparent and trustworthy AI systems (Mozolewski, Robek, and Nalepa, 2024).

To address these challenges, this study conducts a systematic review of 107 empirically validated works published between 2015 and 2025 on explainable fraud detection (XFD) within cyber defense contexts. Using bibliometric mapping, thematic analysis, and technical synthesis, the review examines five key areas critical to cyber warfare resilience: (i) the evolution of XFD research, (ii) ML/DL algorithm foundations, (iii) integration of XAI techniques, (iv) evaluation practices related to explanation quality, and (v) the impact of datasets on transparency, generalizability, and operational trust. Based on these insights, the study proposes a Three-Pillar Framework that conceptualizes explainability not just as a technical feature but as a fundamental socio-technical capability for secure, auditable, and resilient AI systems in financial and insider-threat environments.

The following research questions guide the investigation:

- What patterns of scholarly output, co-authorship clustering, and thematic co-occurrence define the intellectual structure of XFD research between 2015 and 2025?
- Which ML/DL models dominate XFD, and how have their usage and hybridization evolved in response to adversarial and cyber warfare–driven requirements?
- What XAI techniques are most frequently employed, and how do they support transparency, trust, and regulatory or forensic accountability in cyber defense?
- Which evaluation metrics and datasets are prioritized, and how do they shape interpretability, fairness, and real-world operational effectiveness of XFD models?

Section 2 details the methodological approach; Section 3 shows the analytical results; Section 4 discusses these findings in the context of cybersecurity and cyber warfare; and Section 5 highlights research gaps and suggests future directions for creating transparent and resilient AI-powered fraud detection systems.

2. Methodology

This study uses a structured approach to explore how XAI techniques are integrated into the development of transparent financial fraud detection models. The methodology follows a clear protocol emphasizing transparency, reproducibility, and rigor. The process includes four main stages: database selection, search query formulation, screening and eligibility assessment, and bibliometric–thematic analysis.

2.1 Data Sources and Search Strategy

Three reputable databases, Scopus, Web of Science (WoS), and Google Scholar (Gs), were used to ensure comprehensive coverage of peer-reviewed research. The search was limited to the period 2015–2025, reflecting the emergence and rapid evolution of XAI research following the Defense Advanced Research Projects Agency's (DARPA's) formal introduction of the concept in 2016 (Gunning and Aha, 2019) and its subsequent application in financial analytics. The search query combined controlled terms and Boolean operators ("AND," "OR") to capture variations in explainability and fraud detection terminology.

("explainable AI" OR "XAI" OR "model interpretability" OR "model transparency" OR "explainability") AND ("fraud detection" OR "financial fraud" OR "transaction fraud" OR "insurance fraud" OR "banking fraud")

The query was used on article titles, abstracts, and keywords. The initial search returned 575 publications across all databases (Scopus: 374, WoS: 132, Gs: 69).

2.2 Inclusion, Exclusion, and Data Cleaning

Publications were included if they: (i) were peer-reviewed journal articles, conference papers, or book chapters; (ii) were written in English; (iii) examined XAI techniques applied to fraud detection; and (iv) provided empirical validation through experiments or performance evaluation. Studies that were non-peer-reviewed, conceptual-only, or unrelated to fraud detection were excluded. Fifty-three (53) duplicate records were removed using DOI matching and title normalization, followed by manual checks for accuracy. Figure 1 illustrates the screening process.

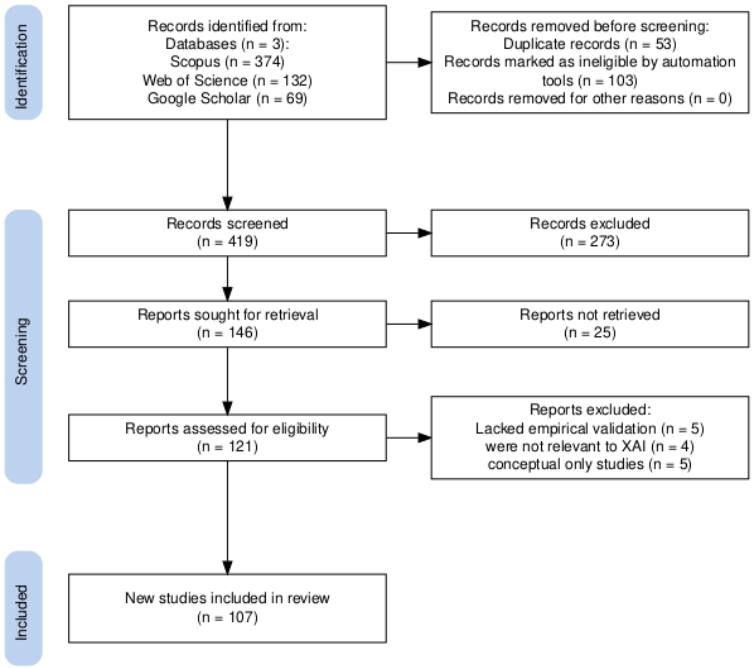


Figure 1: Flowchart for the identification and screening of the dataset

2.3 Screening and Dataset Formation

To ensure that only practical and reproducible studies were included, each publication was screened for empirical validation - that is, the use of real or simulated data rather than purely conceptual or theoretical analysis. Titles and abstracts of all 419 records were reviewed, leading to the exclusion of 273 studies without empirical evidence and 25 subscription-restricted papers. The remaining 121 publications were assessed in full. Fourteen were further excluded: five lacked empirical validation, four were unrelated to XAI, and five were conceptual-only. This process resulted in a final dataset of 107 empirically validated studies (Scopus: 90, WoS: 10, Gs: 7), representing the most relevant and robust contributions.

2.4 Study Quality Assessment

All 107 studies were assessed using the Mixed Methods Appraisal Tool (Nha HONG et al., 2018), which evaluates five criteria: clarity of research questions, suitability of data collection, validity of findings, analytical rigor, and relevance. Each study was rated as High (Yes), Moderate (Partial), or Low (No) quality. Low-quality studies were retained only when they offered unique methodological or technical value.

2.5 Data Extraction

A structured Excel-based form was used to capture key attributes from each study. These include publication details (year, venue, authors, country), fraud domain and dataset used, ML/DL model(s) applied, XAI technique(s) implemented, evaluation metrics used, key findings and contributions, and limitations and future recommendations

2.6 Data Synthesis and Analysis

A mixed bibliometric–thematic approach was employed.

- Bibliometric analysis (using VOS viewer v1.6.20) identified publication trends, collaboration networks, and keyword co-occurrences.
- Thematic analysis categorized studies by intellectual structure, model foundations, XAI adoption, evaluation metrics, and dataset influence on transparency.

A minimum keyword frequency of five was used, with clustering at the default resolution (1.0). Co-authorship networks were also generated to map collaboration patterns.

3. Analysis of the Results

3.1 Intellectual Structure and Collaboration Patterns in XFD Research

Figure 2 shows consistent growth in research on explainable fraud detection (XFD) from 2019 to 2025, with publication volumes accelerating sharply from 2021 onward. Early work focused on basic model comparisons, while studies from 2023–2025 increasingly explored attention mechanisms and intrinsic interpretability. The surge in 2024–2025 signals a shift toward integrating transparency into mainstream fraud analytics, reflecting the sector's heightened emphasis on algorithmic accountability.

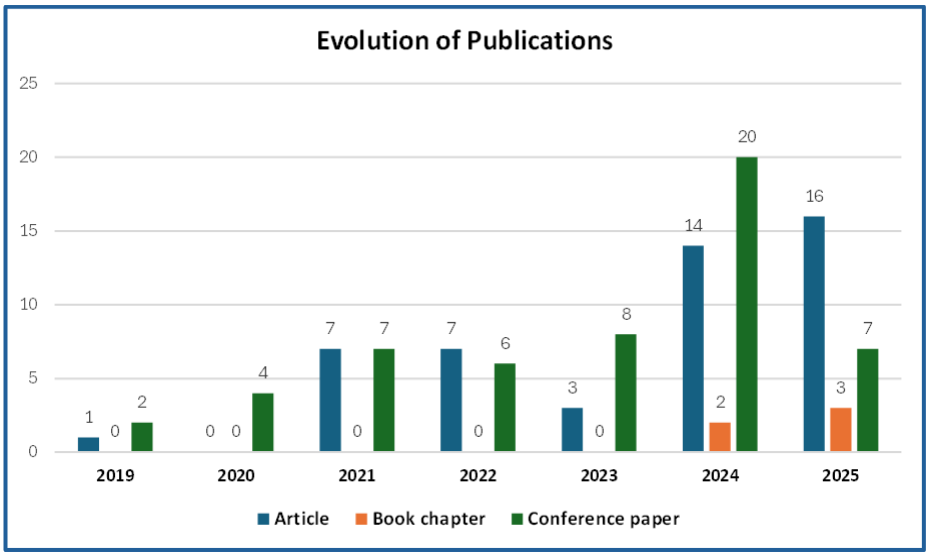


Figure 2: Evolution of Publications

Keyword co-occurrence patterns highlight four thematic clusters: (i) model-agnostic XAI (SHAP, LIME) applied to tabular fraud data; (ii) ML/DL methods for performance–transparency balancing; (iii) privacy-preserving approaches such as federated learning; and (iv) neural architectures leveraging attention and GNNs for relational reasoning. The convergence of "fraud detection," "anomaly detection," "deep learning," and "SHAP" indicates a maturing consensus that explainability must evolve alongside model complexity.

Intellectual activity in XFD reflects growing recognition that transparency, trust, and regulatory alignment must be embedded into fraud detection systems, not added post-hoc.

3.2 Algorithmic Foundations of ML and DL in Fraud Detection

Across the 107 studies, ML models remain dominant (76 occurrences) due to their compatibility with tabular data and interpretability tools. Random Forest (10 studies), XGBoost (8), Decision Trees (7), and Logistic Regression (6) continue to underpin many explainable pipelines. DL methods appear in 54 studies, led by Neural Networks (39), Autoencoders (7), CNNs (6), and LSTMs (4). Table 1 shows the progressive rise of DL from 2019 to 2025, with significant growth in CNNs and Transformers as computational capacity and sequence-modelling needs increased.

Table 1: Trend of ML and DL Techniques Adoption

TECHNIQUE	ML/DL	2019	2020	2021	2022	2023	2024	2025
Transformer	DL	0	0	0	0	0	1	0
Autoencoder	DL	1	0	2	1	0	0	1
Convolutional Neural Network	DL	0	0	0	1	0	2	3
Long Short-Term Memory (LSTMM)	DL	0	0	0	0	1	2	1
Neural Network	DL	2	0	6	5	2	9	9
Recurrent Neural Network (RNN)	DL	0	0	0	1	0	1	0
Transformer	DL	0	0	0	0	1	1	3
Decision Tree	ML	0	1	2	0	1	5	5
Gradient Boosting	ML	0	1	0	1	0	2	6
K-Nearest Neighbors (KNN)	ML	0	0	2	0	0	0	1
Light Gradient Boosting Machine (LightGBM)	ML	0	0	0	0	1	1	2
Logistic Regression	ML	0	1	1	1	2	3	4
Naive Bayes	ML	0	0	0	0	0	0	2
Random Forest	ML	0	1	2	0	3	7	10
Support Vector Machine SVM)	ML	0	0	2	1	1	3	7
Extreme Gradient Boosting XGBoost	ML	0	0	2	0	0	6	8

A dual-track methodological evolution is observed:

- ML models remain widely used due to their transparency and auditability, key requirements in regulated financial environments.
- DL architectures address the rising complexity of fraud patterns, particularly temporal and behavioral dependencies.

Hybrid ML–DL architectures exemplify this socio-technical balancing. Studies such as Roshan and Zafar (2021) and Sakil et al. (2025) combine deep feature extraction, such as Autoencoders and CNNs, with interpretable classifiers, including Random Forest and SHAP-enhanced ML models. This layered strategy mirrors enterprise information systems, where accuracy and interpretability must coexist to support human analysts, auditors, and compliance officers.

3.3 The Role of XAI in Enhancing Transparency and Trust

XAI functions as the bridge between high-performance analytics and the transparency expectations of financial institutions, regulators, and customers. Across the corpus, SHAP, LIME, and Attention Mechanisms dominate, reflecting both their accessibility and alignment with regulatory guidelines such as GDPR's "right to explanation" (The European Parliament, 2016) and Basel's model governance principles (Financial Stability Institute, 2011).

SHAP (34 studies) leads due to its model-agnostic nature and ability to provide both global and local explanations. LIME (15 studies) remains popular for lightweight interpretability, while attention-based models (16 studies) enable intrinsic transparency in sequential and graph-based fraud detection. Figure 3 shows sharp increases in SHAP and LIME after 2022 as financial institutions moved toward auditable AI.

XAI adoption patterns vary by data type:

- SHAP dominates credit card and banking fraud, where features are structured, and explanations must be audit-ready.
- LIME supports logistic regression and Support Vector Machine (SVM) architectures commonly used in insurance claim fraud.
- Attention mechanisms are used for blockchain, identity fraud, and sequential transaction data, where relational dependencies are crucial.

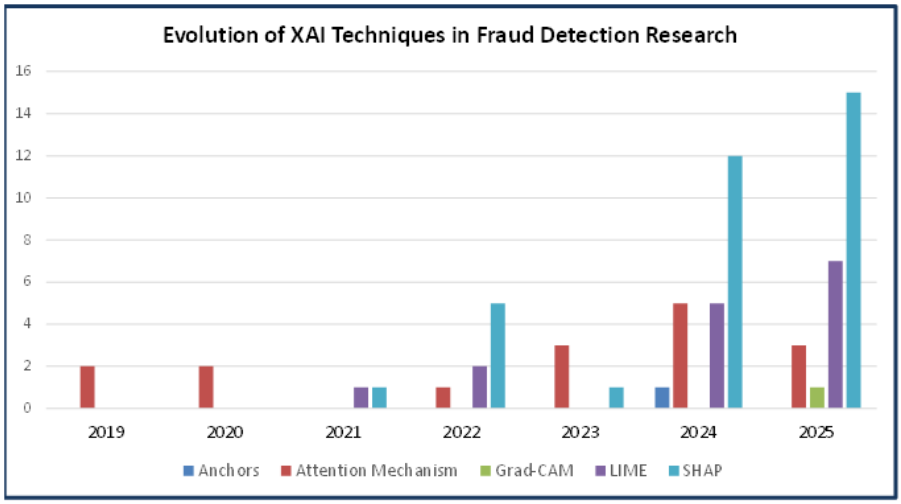


Figure 3: Evolution of XAI Techniques in Fraud Detection Research

However, methods such as counterfactuals, causal explanations, and interactive dashboards remain underused, despite their substantial value for organizational decision-making. Their limited adoption signals an opportunity for further research to enhance end-user comprehension, regulatory reporting, and operational trust.

3.4 Evaluation Metrics and Their Influence on XFD System Design

The literature indicates that Accuracy (48 studies) and Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) (45 studies) dominate XFD evaluations. Their continued use reflects historical habits but fails to capture operational realities in financial institutions, where false positives, false negatives, and the quality of explanations determine trust, compliance, and resource demands.

More operationally meaningful metrics; Recall, Precision, True Positive Rate (TPR), False Positive Rate (FPR), and Mean Squared Error (MSE); remain underused despite their relevance for customer impact, analyst workload, and human–AI decision alignment. This reveals a methodological gap: few studies adopt multi-metric frameworks that integrate predictive performance with explanation fidelity. Such approaches are essential for transparency, governance, auditability, and cross-institutional comparability.

3.5 Dataset Influence on Transparency and Interpretability

Dataset properties significantly affect model transparency. The literature reveals reliance on structured datasets, Financial Datasets (48 studies), Institute of Electrical and Electronics Engineers – Computational Intelligence Society (IEEE-CIS) (36 studies), University of California, Irvine (UCI) (21 studies), and Credit Card Fraud (20 studies), which support interpretable models and effective XAI methods valued by compliance and risk teams.

In contrast, blockchain, mobile money, smart contracts, and insurance datasets are rarely used due to access, standardization, and proprietary constraints, limiting methodological generalizability to emerging digital-finance ecosystems.

Hybrid approaches are gaining traction by integrating structured data with unstructured or relational sources, for example, by combining attention-based blockchain models with SHAP explanations. The findings underscore the need for broader dataset diversity, more precise documentation, and new benchmarks to enhance the transferability and transparency of XFD systems.

3.6 Methodological Limitations

This review focuses solely on peer-reviewed empirical studies published from 2015 to 2025, potentially excluding conceptual or emerging XAI approaches. The English-only criterion introduces linguistic bias, and keyword-based searches may miss studies using unconventional terminology. Although thematic synthesis maintained coherence, some interpretations remain subjective. The review also prioritizes methodological transparency over performance comparison, and findings should be interpreted within this scope.

4. Discussion of Findings

Many fraud detection studies rely on anonymized or synthetic datasets that lack behavioral richness and fail to capture real-world adversarial patterns. Evaluation often focuses narrowly on metrics like accuracy and AUC, neglecting explanation quality, interpretability, and forensic value. This creates a gap: models may perform well statistically but remain unsuitable for cyber defense, where transparency and accountability are crucial. To address this, we propose the Three-Pillar Framework for Explainable AI, which provides a structured approach to building transparent, operationally trustworthy fraud detection systems. These pillars are:

- *Algorithmic Transparency*: the capacity of ML/DL models to expose decision logic through interpretable architectures or post-hoc explanation tools such as SHAP, LIME, or attention mechanisms.
- *Evaluation Accountability*: the use of multi-metric performance assessment that aligns quantitative accuracy with qualitative interpretability, ensuring fairness, fidelity, and stakeholder trust.
- *Data Diversity and Traceability*: the degree to which model explainability reflects the structure, representativeness, and provenance of underlying datasets, enabling reproducibility and auditability.

Together, these pillars define a socio-technical ecosystem in which transparent machine intelligence supports regulatory oversight, ethical governance, and alignment between model behavior and human understanding. Theoretical trends indicate a shift from treating explainability as an external diagnostic layer to integrating it directly into model design, as evidenced by attention-based models and game-theoretic interaction analyses. This evolution positions interpretability as a design objective rather than a post-hoc requirement and suggests future research should optimize for transparency alongside accuracy and robustness.

At the applied level, financial institutions favor risk-averse, legally defensible methods, such as tree-based ensembles paired with SHAP, that provide clear, auditable explanations. Operational constraints, including data privacy, limited cross-institutional datasets, and regulatory inconsistencies, restrict the deployment of more advanced intrinsic-XAI models, widening the gap between exploratory academic work on Transformers, Graph Neural Networks (GNNs), and attention mechanisms and the practical need for simplicity and transparency in regulated environments. Overall, explainability in financial fraud detection is context-dependent and stakeholder-driven: the usefulness of an explanation is shaped not by abstract completeness but by its relevance to regulators, fraud analysts, and customers. Consequently, XAI evaluation in finance should adopt multidimensional criteria that combine technical performance with human-centered qualities such as comprehensibility, actionability, and regulatory compliance.

4.1 Positioning the Three-Pillar Framework within Existing XAI Taxonomies

The proposed Three-Pillar Framework for Explainable Fraud Detection extends existing XAI taxonomies by integrating algorithmic transparency, evaluation accountability, and data traceability into a unified socio-technical model. While rooted in the DARPA-XAI taxonomy (Gunning and Aha, 2019), which focused primarily on model transparency and user-oriented explanations, the framework advances this view by elevating evaluation and dataset provenance as coequal pillars. This reframes explainability from a technical interface problem into a systemic feedback loop that links model logic, performance validation, and data foundations within a single evaluative ecosystem.

The framework also aligns with and expands the XAI 2.0 perspective by Longo et al. (2024), which emphasizes epistemic transparency, ethical accountability, and stakeholder-centric design. Within this lens, Algorithmic Transparency serves as the epistemic layer, Evaluation Accountability as the ethical–regulatory layer, and Data Diversity and Traceability as the empirical foundation for representative and auditable explanations. Synthesizing DARPA-XAI's technical foundations with XAI 2.0's ethical paradigm, the Three-Pillar model operationalizes both in a form tailored to financial AI systems, where interpretability, compliance, and auditability must converge.

4.2 Cross-Domain Applicability of the Three-Pillar Framework

Although developed for financial fraud detection, the Three-Pillar Framework applies broadly to domains involving high-risk, regulated decision-making. In insider threat detection, Algorithmic Transparency supports interpretable behavioral anomaly models by enabling analysts to justify flagged actions. In healthcare fraud and diagnostic prediction, Evaluation Accountability ensures that explanations meet ethical and clinical standards. Likewise, Data Diversity and Traceability are crucial in cybersecurity and critical infrastructure monitoring, where verifiable data provenance underpins auditing, forensics, and incident response.

Collectively, these mappings demonstrate that the framework offers a scalable, domain-agnostic architecture for trustworthy AI in sectors where transparency, accountability, and auditability are central to operational integrity. Treating explainability as a measurable, reproducible property of model design rather than a purely qualitative add-on, the framework advances the goal of developing human-centered, reliable metrics for transparent intelligent systems.

5. Key Findings and Conclusion

This study identifies prevailing trends, research gaps, and prospective directions to advance transparency, interpretability, and trust in AI-driven decision systems.

5.1 Research Gaps

- Existing explainable fraud and insider-threat detection models rarely use recurrent architectures or latent behavioral representations. As a result, they struggle to model evolving attack behavior.
- The field exhibits heavy reliance on post-hoc, model-agnostic tools such as SHAP and LIME, with minimal exploration of causal, counterfactual, or visual explanation methods like Gradient-weighted Class Activation Mapping (Grad-CAM) and Anchors. This methodological homogeneity restricts interpretability to feature attribution rather than causal reasoning or decision transparency, an essential requirement in regulated financial environments.
- Evaluation practices focus predominantly on Accuracy and ROC–AUC, with limited consideration of domain-sensitive and explanation-centric metrics such as Recall, Precision, False Positive Rate, explanation fidelity, or user satisfaction. This imbalance risks overstating performance and obscuring the operational reliability of models deployed in real-world financial systems.
- Research is heavily concentrated on a few structured and publicly available datasets (IEEE-CIS, UCI Repository, and Credit Card datasets). Emerging ecosystems, blockchain transactions, smart contracts, mobile money, and insurance claims remain underrepresented, limiting model generalizability and contextual validity.

5.2 Conclusions

This paper reviewed the role of Explainable Artificial Intelligence (XAI) in financial and insider fraud detection within cyber warfare and security contexts. The review showed that most existing systems prioritize predictive accuracy over transparency. This imbalance reduces analyst trust, weakens forensic investigation, and complicates regulatory compliance. In high-risk security environments, black-box models hinder effective, accountable decision-making.

To address these challenges, the paper proposed a Three-Pillar Framework for Explainable Fraud Detection. The framework consists of Algorithmic Transparency, Evaluation Accountability, and Data Diversity and Traceability. It treats explainability as a core design requirement rather than a post-hoc feature. The framework provides a structured approach for aligning AI models with operational, regulatory, and human-centered needs in cyber defense settings.

Algorithmic Transparency focuses on making model decisions understandable to analysts. This can be achieved through interpretable models, attention mechanisms, or post-hoc tools such as SHAP and LIME. Evaluation Accountability extends performance assessment beyond accuracy and ROC–AUC. It includes operational metrics and explanation quality to support trust, auditability, and responsible use. Data Diversity and Traceability ensure that datasets are representative, well-documented, and auditable. This supports reproducibility and regulatory review.

A key gap identified in this review is the limited use of temporal and behavioral modeling in explainable fraud detection. Such models are essential for capturing how fraud evolves. Variational Autoencoders (VAEs) can represent latent behavioral patterns, while Bidirectional LSTMs (BiLSTMs) can model sequential dependencies. When applied within the proposed Three-Pillar Framework, these approaches can guide future research on insider financial fraud detection while maintaining interpretability and auditability.

Ethics Declaration: The authors declare that ethical clearance was not required for the research. No human subjects or sensitive data used

AI Declaration: The authors declare that AI tools were not used to create this paper. Grammarly was employed to ensure clarity, grammar, and professional tone in the documentation.

References

Abbas, Z.A., Hilal, Z.M., Jabbar, H.G., 2025. Click Fraud Detection in Online Advertising: A Comparative Study of Machine Learning Models. International Journal of Safety and Security Engineering 15, pp 427–441. https://doi.org/10.18280/IJSSE.150303

Adetumi Adewumi, Somto Emmanuel Ewim, Ngodoo Joy Sam-Bulya, Olajumoke Bolatito Ajani, 2024. Enhancing financial fraud detection using adaptive machine learning models and business analytics. International Journal of Scientific Research Updates 8, pp 012–021. https://doi.org/10.53430/IJSRU.2024.8.2.0054

Agomuo, O.C., Uzoma, A.K., Khan, Z., Otuomasirichi, A.I., Muzamal, J.H., 2025. Transparent AI for Adaptive Fraud Detection. Proceedings of the 2025 19th International Conference on Ubiquitous Information Management and Communication, IMCOM 2025. https://doi.org/10.1109/IMCOM64595.2025.10857433

Ahmad, W., Vashist, A., Sinha, N., Prasad, M., Shrivastava, V., Muzamal, J.H., 2025. Enhancing Transparency and Privacy in Financial Fraud Detection: The Integration of Explainable AI and Federated Learning. Communications in Computer and Information Science 2244 CCIS, pp 139–156. https://doi.org/10.1007/978-3-031-75201-8_10

Al-hchaimi, A.A.J., Alomari, M.F., Muhsen, Y.R., Sulaiman, N. Bin, Ali, S.H., 2024. Explainable Machine Learning for Real-Time Payment Fraud Detection: Building Trustworthy Models to Protect Financial Transactions. Lecture Notes in Networks and Systems 1033 LNNS, pp 1–25. https://doi.org/10.1007/978-3-031-63717-9_1

Aljunaid, S.K., Almheiri, S.J., Dawood, H., Khan, M.A., 2025. Secure and Transparent Banking: Explainable AI-Driven Federated Learning Model for Financial Fraud Detection. Journal of Risk and Financial Management 18. https://doi.org/10.3390/JRFM18040179

Ezzeddine, F., Saad, M., Ayoub, O., Andreoletti, D., Gjoreski, M., Sbeity, I., Langheinrich, M., Giordano, S., 2024. Differential Privacy for Anomaly Detection: Analyzing the Trade-Off Between Privacy and Explainability. Communications in Computer and Information Science 2155 CCIS, pp 294–318. https://doi.org/10.1007/978-3-031-63800-8_15

Financial Stability Institute, 2011. Principles for the Sound Management of Operational Risk (PSMOR). https://www.bis.org/publ/bcbs195.pdf.

Gunning, D., Aha, D.W., 2019. DARPA's explainable artificial intelligence program. AI Mag 40, pp 44–58. https://doi.org/10.1609/AIMAG.V40I2.2850

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. Del, Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., Stumpf, S., 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion 106, 102301. https://doi.org/10.1016/J.INFFUS.2024.102301

Mozolewski, M., Bobek, S., Nalepa, G.J., 2024. Visual Explanations and Perturbation-Based Fidelity Metrics for Feature-Based Models. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 14835 LNCS, pp 294–309. https://doi.org/10.1007/978-3-031-63772-8_27

Nha HONG, Q., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., Rousseau, M.-C., Vedel, I., n.d. MIXED METHODS APPRAISAL TOOL (MMAT) VERSION 2018 User guide.

Ofori, H.K., Bell-Dzide, K., Brown-Acquaye, W.L., Lempogo, F., Frimpong, S.O., Agbehadji, I.E., Millham, R.C., 2025. Application of Machine Learning and Deep Learning Techniques for Enhanced Insider Threat Detection in Cybersecurity: Bibliometric Review. Symmetry 2025, Vol. 17, Page 1704 17, 1704. https://doi.org/10.3390/SYM17101704

Pennella, L., Pinelli, F., Galletta, L., 2025. X-SPIDE: An eXplainable Machine Learning Pipeline for Detecting Smart Ponzi Contracts in Ethereum. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3569565

Praveenraj, D.D.W., Victor, M., Vennila, C., Alawadi, A.H., Diyora, P., Vasudevan, N., Avudaiappan, T., 2023. Exploring Explainable Artificial Intelligence for Transparent Decision Making. E3S Web of Conferences 399. https://doi.org/10.1051/E3SCONF/202339904030

Qin, Z., Liu, Y., He, Q., Ao, X., 2022. Explainable Graph-based Fraud Detection via Neural Meta-graph Search. International Conference on Information and Knowledge Management, Proceedings pp 4414–4418. https://doi.org/10.1145/3511808.3557598

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 13-17-August-2016, pp 1135–1144. https://doi.org/10.1145/2939672.2939778

Roshan, K., Zafar, A., 2021. Utilizing Xai Technique to Improve Autoencoder Based Model for Computer Network Anomaly Detection with Shapley Additive Explanation(SHAP). International Journal of Computer Networks and Communications 13, pp 109–128. https://doi.org/10.5121/IJCNC.2021.13607

Sakil, M.B.H., Hasan, M.A., Mozumder, M.S.A., Hasan, M.R., Opee, S.A., Mridha, M.F., Aung, Z., 2025. Enhancing Medicare Fraud Detection with a CNN-Transformer-XGBoost Framework and Explainable AI. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3562577

The European Parliament, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonized rules on artificial intelligence. Office of the European Union L-, Publications.

The European Parliament, 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

Tian, Y., 2025. Financial Fraud Detection Based on an XGBoost and LSTM Fusion Model: A Comparative Study on the Enhancement of Time-Series Features. Advances in Economics, Management and Political Sciences 170, pp 101–111. https://doi.org/10.54254/2754-1169/2025.LH23993

Vivek, Y., Ravi, V., Mane, A., Naidu, L.R., 2025. Explainable One Class Classification for ATM Fraud Detection. International Conference on Communication Systems and Networks, COMSNETS pp 114–119. https://doi.org/10.1109/COMSNETS63942.2025.10885737

Zhang, K., Wang, H., Chen, M., Chen, X., Liu, L., Geng, Q., Zhou, Y., 2025. Leveraging machine learning to proactively identify phishing campaigns before they strike. J Big Data 12. https://doi.org/10.1186/S40537-025-01174-X

Zhu, W., Chen, Z., 2024. An Intelligent Financial Fraud Detection Model Using Knowledge Graph-Integrated Deep Neural Network. Journal of Circuits, Systems and Computers 33. https://doi.org/10.1142/S0218126624502670

Zhu, W., Zhang, C., Li, J., Wang, Z., 2025. An Intelligent Financial Fraud Detection Model Based on Dilated Convolution and Generative Adversarial Network. Journal of Circuits, Systems and Computers 34. https://doi.org/10.1142/S0218126625501701

Cyber-Physical Incident Attribution in UAV/Rail Attacks

Isaac Ojeh[1], Xavier Palmer[2] and Lucas Potter[2]
[1]MorphHats InfoSecure, Waterloo, Canada
[2]BiosView Labs, Dayton, Ohio

morpheus@morphhats.com
Biosview1@proton.me

Abstract: As unmanned aerial vehicles (UAVs) and smart rail systems become increasingly integrated into critical logistics infrastructure, they also present new surfaces for hybrid cyber-physical attacks. Coordinated adversarial actions such as cyber intrusions that manipulate physical trajectories, sensor spoofing, or disruptions to control systems pose significant challenges for real-time detection and post-incident analysis. Effective attribution of such incidents is crucial not only for identifying responsible parties but also for enhancing resilience and enabling coordinated defense responses across infrastructure operators, governments, and private stakeholders. This paper examines the problem of cyber-physical incident attribution in the context of combined UAV/rail attacks, where attack vectors may span networked systems, edge devices, and physical actuators. We propose a layered attribution framework that fuses telemetry from cyber logs, UAV flight data, rail signaling systems, and environmental sensors to reconstruct the sequence and origin of coordinated attacks. The system leverages graph-based causality analysis, trust scoring mechanisms, and cross-domain forensic correlation to associate anomalies with likely sources and attack pathways. Our approach combines both deterministic rules and machine learning models trained on simulated and real-world incident data to balance explainability and adaptive intelligence. This paper, overall, adopts an exploratory perspective, examining foundational challenges and design trade-offs involved in attributing cyber-physical incidents within a multimodal UAV/rail logistics environment. Rather than proposing a finalized solution, the work seeks to identify key data-fusion requirements, threat-modeling gaps, and policy implications to inform future technical and legal frameworks for attribution. Preliminary results from simulated hybrid attacks using a digital twin environment show promising attribution accuracy, particularly when incorporating temporal patterns and system interdependencies. However, limitations in sensor coverage and adversarial evasion tactics underscore the need for multi-source trust validation and international collaboration in standardizing attribution protocols. Ultimately, this research aims to lay groundwork for a scalable, context-aware attribution system that can support accountability, deterrence, and rapid response in the evolving landscape of autonomous transportation.

Keywords: Cyber physical security, UAV attacks, Rail infrastructure, Incident attribution, Forensic analysis, Hybrid threats, Digital twins, Critical infrastructure defense

1. Introduction

Some modern logistics can benefit from the integration of cyber-physical systems such as UAVs and smart railway networks; this can enhance efficiency but can also carry the risk of expanding attack surfaces. Coordinated cyber-physical incidents, such as a cyber intrusion into rail signaling combined with drone hijacking, are hard to detect and attribute (Longo et al. 2025). Attribution, or identifying responsibility, can be made more complex in these multi-domain scenarios. Traditional cybersecurity already faces hurdles, and hybrid UAV-rail attacks amplify this due to disparate systems (IT, wireless, mechanical) and unclear intrusion signs (Hartmann and Giles, 2016). Effective attribution is crucial for accountability, deterrence and improved incident response (Rid and Buchanan 2015). Despite its importance, cyber-physical incident attribution is underdeveloped. Existing railway cybersecurity and UAV security research focus on prevention, not sufficiently on post-incident attribution (Ibadah et al. 2024; Soderi et al. 2023). This gap stems from siloed monitoring, incompatible data, and a lack of cross-domain forensic standards. This paper aims to address that through proposing a layered attribution framework for UAV-rail incidents. It fuses data from various sources (network logs, UAV telemetry, sensors, surveillance) to build an event timeline, inferring attack pathways and entities through cause-and-effect analysis. Techniques include graph-based causality modeling and machine learning for pattern recognition (Qiu et al. 2024). Our exploratory work outlines the framework's architecture, discusses challenges (data integrity, real-time analysis, explainability), and presents a preliminary evaluation using a digital twin. Results show that correlating cyber and physical telemetry significantly improves attribution accuracy, identifying malicious events across domains where isolated analysis failed. Limitations include sensor blind spots and adversarial data injection. This research aims to build foundational, context-aware attribution systems for critical autonomous transportation. As railways integrate IoT and drones become ubiquitous, the security community must preempt attackers exploiting these convergences. Lessons learned will inform technical designs and foster collaborations for collective defense. The following section reviews related work and threats in UAV/rail environments.

1.1 Definition of Cyber-Physical Incident Attribution

In this paper, cyber-physical incident attribution is defined as the structured process of identifying the likely origin, pathway, and responsible entities behind a coordinated cyber-physical attack by correlating technical evidence across digital and physical domains. This definition is intentionally scoped to technical and operational attribution rather than legal adjudication. Attribution in cybersecurity literature often spans multiple contexts. In technical cybersecurity, attribution refers to identifying the infrastructure, tools, and tactics used in an attack through forensic analysis and threat intelligence (Rid and Buchanan, 2015). In criminal law, attribution establishes an individual or group's responsibility for prosecution. In contrast, international law determines whether a state is accountable for wrongful acts committed directly or through proxies. While these legal interpretations are critical, they fall outside the scope of this study. This work focuses specifically on technical attribution within cyber-physical systems, aiming to reconstruct attack sequences, correlate cross-domain evidence, and support investigative decision-making. Legal responsibility and political assessment are treated as downstream processes that may build upon the technical attribution outputs generated by the proposed framework.

2. Background and Related Work

Cyber-physical attacks, such as Stuxnet (2010), show that digital actions can have physical consequences. In rail, incidents range from malware to ransomware, exposing vulnerabilities (Karnouskos, 2011). UAVs pose new threats, with drones interfering with airports or being hijacked. Both rail and UAV domains face compounding cyber-physical risks. Modern rail systems are industrial control systems (ICS). Securing rail ICS is crucial due to safety and economic implications. While standards exist, gaps remain in covering legacy equipment and integrating physical safety with cyber threat detection. Traditional safety systems assume benign failures, necessitating new approaches that combine operational and cybersecurity data for incident analysis. UAVs are flying cyber-physical systems. UAV platforms share architectural similarities with industrial control systems, making them susceptible to comparable cyber threats. Research has identified threats such as GPS spoofing, command-link hijacking, and sensor manipulation in UAV platforms (Kim et al. 2021). Many UAV platforms share ICS characteristics, enabling the adaptation of industrial cybersecurity techniques. However, UAVs introduce unique challenges as they are mobile, operate in unregulated airspace, and can reach sensitive locations. Attribution in cyberattacks is a studied problem that involves tracing activity through digital forensics. Cyber-physical incident attribution is a newer approach that extends digital forensics to include physical causal reasoning. Frameworks are emerging that correlate cyber and physical event patterns to distinguish equipment failures from cyberattacks. Cross-domain correlation techniques have been shown to distinguish cyberattacks from physical faults in power systems. Cross-domain correlation of cyber and physical events has been shown to improve the distinction between system faults and deliberate cyberattacks (Semertzis et al. 2024). Digital forensics readiness is critical, and methods such as digital twins improve accident investigation accuracy. Maintaining extensive logs and synchronized, trustworthy data is essential for post-incident analysis. The convergence of UAV and rail systems is practical, with drones used for inspection and security. For instance, Sensonic uses fiber optic sensors to detect anomalies and dispatch drones for visual confirmation. This integration, while improving response, also creates potential attack vectors that adversaries could exploit to manipulate sensors and hijack drones. Commercial rail operators are already integrating drones with track monitoring systems, increasing both resilience and attack surface. Rail operators have begun integrating drones with track monitoring and security systems, expanding both operational resilience and the attack surface (Pyke 2025). While no sufficient incidents of this nature have been reported, the individual components are vulnerable, suggesting multi-vector attacks are a plausible future threat. In conclusion, cyber and physical aspects of critical infrastructure are increasingly interdependent. Our work aims to bridge the gap in investigative and attribution methods for UAV/rail contexts, drawing on insights from ICS security, drone forensics, and multi-stage attack analysis.

2.1 Indicators of Compromise and Indicators of Attack

Cyber-physical attribution relies on multiple classes of indicators. Traditional cyber forensics emphasize Indicators of Compromise (IoCs), which are technical artifacts such as malicious IP addresses, file hashes, or anomalous log entries that indicate a system has been compromised (Haber and Rolls, 2019; Haber and Rolls, 2024). IoCs are typically reactive and focus on identifying what was affected. In contrast, Indicators of Attack (IoAs) are behavioral and contextual indicators that describe how and why an attack is unfolding (Hwang and Kim, 2023; Yi and Kim, 2024). IoAs capture patterns such as coordinated timing across systems, repeated manipulation of control logic, or consistent sequencing of cyber and physical actions. These indicators are

particularly relevant for attribution, as they provide insight into adversary intent and operational design rather than isolated technical artifacts. The framework proposed in this paper integrates both IoCs and IoAs. IoCs support detection and forensic validation, while IoAs inform causal graph construction and hypothesis generation, enabling attribution of coordinated cyber-physical attacks beyond simple compromise detection.

2.2 Hybrid Cyber-Physical Attacks in UAV and Rail Systems

Hybrid cyber-physical attacks are defined here as coordinated adversarial actions in which cyber operations are deliberately used to induce, amplify, or conceal physical effects (Simons et al, 2020; Lian et al, 2024). In UAV and rail environments, such attacks often involve multiple stages, including digital intrusion, control system manipulation, sensor spoofing, and physical actuation. Examples include unauthorized modification of rail signaling logic combined with UAV-based reconnaissance or obstruction, GPS spoofing used to redirect drones while masking physical intrusions, or denial-of-service attacks that delay operator response during a physical sabotage event. These attacks differ from purely cyber or purely physical incidents in that their effects emerge from the interaction between digital commands and physical system behavior. Understanding these hybrid attack structures is essential for attribution, as isolated analysis of either cyber logs or physical anomalies is often insufficient to reconstruct the whole attack sequence.

3. Layered Attribution Framework

3.1 Design Objectives and Overview

The primary goal of the proposed framework is to reconstruct the sequence of events in a cyber-physical attack and identify the responsible entities (be it specific attackers, malware, or compromised devices) with high confidence. To achieve this, the framework is designed with several objectives in mind. To start, a comprehensive framework is essential, capturing data from all relevant layers, including network traffic, system logs, application-level events, physical sensor outputs, and operator actions, as no single vantage point suffices for complex incidents. For real-time or near-real-time responses, timeliness is crucial, requiring the framework to process and correlate events as they happen and to raise attribution hypotheses even before an incident entirely unfolds; however, offline forensic analysis is also supported for thorough post-incident investigations. The framework must also assess the trustworthiness and integrity of data sources, recognize when an attacker might have tampered with logs or sensor readings, and incorporate redundant or independent data sources to help validate events. Furthermore, the results of the analysis of the attribution should be explainable and interpretable by humans, especially in critical infrastructure, where accountability is key. They should provide a chain of supporting evidence or reasoning rather than just an actor's attribution. These principles lay the groundwork for a system that can effectively adapt to evolving threats.

Table 1: Layered Attribution Framework for UAV and Rail Incidents

Layer	Purpose	Example Data Sources
Data Collection	Capture cyber and physical telemetry	Rail signaling logs, UAV flight data, network flow logs
Event Processing	Detect anomalies and normalize data	Safety rule violations, ML-based anomaly alerts
Causality Analysis	Correlating events across domains	Temporal and causal event graphs
Attribution Inference	Generate and rank attack hypotheses	IoC and IoA correlation, trust-weighted reasoning
Investigation Interface	Support analyst review and reporting	Timelines, evidence chains, hypothesis summaries

Our proposed layered architecture, depicted in Figure 1, includes several modules designed to achieve our objectives. At the foundational data collection layer, interfaces gather information from various sources. These sources include UAV telemetry streams, such as flight path, controller inputs, GPS data, and payload status, as well as rail system logs, which encompass signaling commands, track sensor triggers, and train control system data. Additionally, network monitors collect traffic from relevant networks, potentially including attacker command-and-control communications. Physical surveillance, including CCTV footage and distributed-acoustic-sensing fiber alerts, also contributes to the data. Finally, environmental data, such as GPS timing signals and weather conditions, is crucial for ruling out natural causes, such as solar storms or GPS errors. The system processes data in several layers. First, Event Processing and Anomaly Detection analyze data streams using rule-based and algorithmic methods to flag unusual events, applying domain-specific knowledge (Ibadah et al. 2024). Next, the Causality Graph Construction module links timestamped events from various sources into a graph where nodes are events/states and edges represent causal/temporal relationships. This approach, inspired by attack graph methodologies, establishes edges through deterministic logic or learned patterns (Qiu et al. 2024).

Each graph element carries a trust score that reflects sensor reliability. If a sensor is compromised, its data receives lower confidence unless corroborated. The Attribution Inference Engine then analyzes event clusters and matches them to known attack signatures or threat actors. This involves hypothesis generation, potentially using knowledge bases such as MITRE ATT&CK to label tactics, techniques, and procedures (TTPs), which can hint at attribution (Strom et al., 2018). Finally, an Investigation Interface provides human analysts with visualizations, causal graphs, and narrative explanations, linking evidence to conclusions. It also allows analysts to verify data for legal defensibility.

3.2 Graph-Based Causality Analysis

Our framework's foundation lies in the construction of an event-causality graph, an approach supported by previous research that has used graph models to untangle intricate incident chains (Casey et al. 2015; Qiu et al. 2024). Within this graph, we delineate several node types. These include Cyber Event Nodes, such as "Firewall log entry: port scan detected" or "PLC command issued: switch track," and Physical Event Nodes, such as "Drone altitude change" or "Train emergency brake activated." Additionally, Alert/Anomaly Nodes represent triggers from detection algorithms, like an "Intrusion Detection System alert." At the same time, State Nodes capture relevant system states or conditions, for instance, "Train A was in section X at time T." Understanding these node types is crucial for accurately mapping the sequence of events during a cyber-physical incident. Edges show causal or temporal links. Causal links indicate direct precipitation, while temporal edges suggest ordering and potential influence. Machine learning also identifies correlated events, like network latency anomalies linked to sensor faults. To manage incident complexity, we layer graphs, maintaining subgraphs per system (e.g., UAV, rail) with cross-links. This allows analysts to view simplified slices of an incident (e.g., drone vs. train events) and their convergence in the larger picture, like network attack graph layers. Graph-based analysis systematically evaluates attack paths and assesses the likelihood of each. We use metrics like out-degree and betweenness to identify critical pivot points, similar to Qiu et al. (2024). A single malicious packet causing downstream failures would appear as a high-influence node, highlighting the attacker's pivotal action.

3.3 Trust Scoring Mechanism

Our framework's reliance on fusing data from multiple sources necessitates considering the potential compromise or unreliability of some sources. For instance, if an attacker gains administrative access to a control server, the server's logs could be manipulated to conceal their activities. Unquestioning trust in all logs could lead to erroneous conclusions, such as attributing a malfunction when, in fact, a malicious command was deliberately deleted from the log. We propose a straightforward trust-scoring mechanism: each data source, and, by extension, every event originating from it, is assigned a numerical trust level. This assignment can be static, based on prior assessments of the source's security hardening, or dynamic, depending on whether the system exhibits normal behavior during an incident. For example, an independent GPS sensor on a drone might be deemed more trustworthy than the drone's internal GPS if GPS spoofing is suspected, as a tamper-proof independent sensor could detect discrepancies. Similarly, if a rail control center is isolated and highly secured, its logs might be assigned to high trust; however, if abnormal behavior or an intrusion alert is triggered during an incident, confidence in its data would consequently diminish. In practice, trust scores influence analysis by weighing the edges in the causality graph. Trust-weighted sensor fusion has been proposed as a mechanism for improving reliability in cyber-physical systems. Trust-weighted sensor data fusion improves reliability in cyber-physical attribution by accounting for compromised or unreliable data sources (Yadav et al. 2023). An event chain dependent on low-trust data will be regarded with skepticism unless corroborated. The inference engine can present alternative hypotheses: one assuming data accuracy, another assuming data fabrication. By comparing these, investigators can determine whether an attacker may have concealed their tracks. In essence, our framework endeavors to achieve forensic robustness by avoiding exclusive reliance on a single data source.

3.4 Machine Learning Components

While rule-based detection and expert logic are fundamental for explainability, machine learning (ML) offers the adaptability and pattern recognition necessary to address emerging threats. We apply ML in two key areas: anomaly detection and attack classification/attribution. For anomaly detection, we deploy unsupervised or semi-supervised learning algorithms, such as clustering, autoencoders, or one-class SVMs, to model the normal behavior of UAV operations and rail systems, subsequently flagging outliers. For instance, a deep learning model could learn the typical correlation between train speed profiles and signaling commands, where any deviation might suggest interference. Regarding attack classification/attribution, we train supervised models using simulated incident data to categorize attack types and even link them to known attacker profiles. For example, if a particular threat actor consistently employs a specific sequence of actions (e.g., jamming GPS followed by

injecting malware), a classifier can identify this pattern in incoming data and label it accordingly (e.g., "Pattern resembles Attack Group A's known operations"). This information can then contribute to attribution hypotheses. It is crucial to thoroughly test these ML models in the simulation environment to prevent overfitting to expected patterns and to ensure they can handle real-world noise and variance. Furthermore, by integrating deterministic rules with ML outputs, we aim to prevent "black-box" decision-making. ML suggestions are evaluated alongside logical rules (e.g., "if a valid operator command was never received, the train should not have accelerated; if it did, that's a strong indication of malicious control"). Essentially, we do not permit the system to make an attribution claim based solely on an ML output; ML assists in prioritizing and guiding the search for supporting evidence, but final conclusions require corroboration through the causal graph.

3.5 Data Fusion and Cross-Domain Correlation

Our framework's core is data fusion, correlating events across domains to uncover insights that single-domain analysis misses. For example, simultaneous rail signal errors and drone connection loss, while seemingly isolated, strongly suggest a linked cause like electromagnetic interference or a cyberattack. Our system automatically identifies these temporal and contextual alignments. To achieve this, all data is time-stamped with a common reference clock (NTP/GPS for UAV and rail systems) and normalized into a unified schema (time, location, severity, etc.). Distributed system techniques handle clock drift. Cross-domain correlation also identifies higher-level patterns. Multiple anomalies across subsystems in a short period often indicate a broader attack. Our heuristic: "if X or more subsystems report significant anomalies within Y minutes, suspect a coordinated attack," mirrors observations in past power grid incidents. In summary, this layered framework systematically transforms raw data into a coherent attack narrative and attribution.

4. Challenges and Trade-Offs in Attribution

Designing an attribution system for cyber-physical incidents presents several inherent challenges, including managing high data volumes while ensuring real-time analysis. Our framework addresses this by filtering relevant or abnormal data from trains and UAVs to reduce load; however, this risks false negatives, which we mitigate by sampling normal data and updating ML models. Another challenge involves balancing deterministic rules and machine learning, where we prioritize ML's strong indication of an attack over rules, accepting more false positives for critical infrastructure safety. To provide human-understandable explanations for ML outputs despite their complexity, we utilize causal graphs to link anomalies and attack classifications to evidence, such as sensor deviations or code signatures, ensuring ML informs without replacing evidence. Furthermore, we account for sensor and coverage gaps by simulating data gaps and flagging "insufficient data" in the analysis, acknowledging uncertainty while planning future mitigation through redundant sensors and data sharing. Although false flags and deception are not entirely solvable, multi-source correlation helps expose inconsistencies in fabricated attack scenarios, and we err on the side of caution, suspecting attacks if other anomalies are detected concurrently. Finally, privacy and legal considerations are addressed by ensuring data integrity through timestamping, cryptographic signing, and tamper-evident storage, and by filtering or anonymizing personally identifiable information; compliance also necessitates transparent governance on data usage. Despite these challenges, integrated attribution demonstrates promising benefits, as detailed in our simulation-based evaluation.

5. Evaluation in a Simulated Environment

We evaluated our framework using a digital twin simulation of a logistics corridor with a railway and UAVs. Digital twin approaches have also been applied to UAV forensic analysis, improving post-incident reconstruction accuracy. The framework was evaluated through a digital twin simulation that combined a rail control system and UAV operations (Almusayli et al. 2024). The simulation combined an open-source train control simulator with ROS 2 and Gazebo for UAVs, mimicking realistic signaling and drone operations, including communication networks. We designed attack scenarios based on threat models. One, "Coordinated Derailment Attempt," involved an attacker penetrating the rail network, changing a track switch to derail a train, jamming a drone's GPS, hijacking a surveillance drone to drop a payload onto the tracks, and launching a denial-of-service attack. The attacker also attempted to erase system logs. Our attribution framework, connected in real time, ingested data such as train speed, signal states, drone telemetry, and network logs. It successfully detected anomalies, correlating the unauthorized track switch command with erratic drone GPS and flight deviation, flagging a "Possible coordinated attack: rail switch manipulation and drone deviation concurrently observed. Likely sabotage attempt." It identified log tampering and, using its knowledge base, categorized the incident as a deliberate cyber-physical attack. In 9 out of 10 runs, the framework correctly associated the rail and UAV

anomalies. Another "False Flag" scenario tested the framework's ability to identify technical events while providing two possible hypotheses (malware signature vs. behavior pattern) to the analyst, aligning with our design principle of aiding human judgment. Performance metrics showed the system processed around 500 events per second, sufficient for simulated scenarios. Limitations included the variability of real-world systems and the challenge of tuning anomaly detectors. Cascade effects from a train derailment also temporarily overwhelmed the causal graph, highlighting the need for better abstraction. Nonetheless, the simulation confirmed the approach's viability, significantly improving situational awareness by integrating data that would traditionally be siloed. The use of a digital twin environment enabled controlled experimentation with hybrid attack scenarios, allowing repeatable testing of attribution accuracy under varying conditions without risk to operational infrastructure.

6. Discussion, Limitations, and Future Works

Our framework shows promise for cyber-physical attribution in UAV and rail incidents. Robust attribution requires multidisciplinary collaboration, including IT security, rail engineers, drone operators, and forensic analysts. Organizations must foster cross-domain teams that integrate IT, OT, and UAV experts, especially as they confront an increasingly multidisciplinary cybersecurity landscape (Abdo and Hossain, 2025). Standardized data formats and sharing protocols are crucial for scalable attribution. Policymakers should mandate greater data exchanges, addressing liability and ownership issues. International cooperation, through bodies such as ICAO, is vital for addressing cross-border threats (Minnaar, 2022; Weber, 2025). Adversaries can be expected to evade attribution, necessitating continuous refinement, adversarial testing, and integration of real-time threat intelligence. Our methodology, which fuses multi-source telemetry and causal graphs, applies to diverse cyber-physical systems, such as autonomous trucks and maritime drones. Ethical and privacy concerns demand strict policies and oversight to prevent misuse of surveillance data. Transparency and independent audits remain essential. Investigations are ethically imperative, supported by our framework's data fidelity and analytical transparency. A limitation is the need for more straightforward explanations for non-technical stakeholders. A user-friendly explanation module could generate plain-language incident reports, translating technical insights into actionable intelligence for legal and policy purposes. Implementing such a system requires significant investment, but it is essential. Prevention is not foolproof; however, attribution aids response, deter attackers, and strengthens defenses. A phased rollout focused on critical operations is advisable. Attribution must integrate seamlessly with incident response, providing intelligence for containment and remediation and supporting law enforcement. While not a complete solution, our research is a crucial step towards demystifying cyber-physical attacks. By clarifying the unfolding of attacks, in their many forms and combinations, we empower defenders and responders, making sophisticated attribution a cornerstone of future safety and security. Future work involves real-world testing of the framework with industry partners using benign scenarios or historical data, alongside enhanced automation to streamline investigations by automatically querying related systems for corroborating data. Within this span of efforts, a focus on protecting infrastructure and workflows more closely tied to the bioeconomy across land and water is of interest due to logistical necessities with some biological handling and processing that have increasingly involved digital means (Potter and Palmer, 2023; Barnett et al., 2024). Furthermore, integrating with legal frameworks through collaboration with policymakers on data sharing and liability for cyber-physical incidents is crucial, as is developing predictive defenses by adapting elements for early warning or prediction through the detection of attack precursors. Finally, understanding human factors, including user interaction, interface design, and training, is essential for effective use during incidents.

7. Conclusion

Autonomous systems like UAVs and smart railways offer immense benefits but introduce new security vulnerabilities spanning cyber and physical domains. This paper addressed the attribution of incidents across these domains, which is crucial for accountability and defense in logistics and transportation. Using graph-based causality, it linked malicious cyber actions (e.g., network intrusion) to physical consequences (e.g., train malfunction). Trust scoring and combined deterministic and machine-learning methods handle uncertainty. Our simulated UAV/rail implementation showed the system could identify coordinated malicious actions that would otherwise appear benign. Key contributions included highlighting data-fusion needs for cross-domain attribution, identifying system gaps (e.g., the lack of rail control/UAV telemetry integration), and providing initial evidence that unified analysis can detect complex attack sequences. We also discussed design trade-offs and made recommendations. In summary, attributing cyber-physical incidents in multimodal environments is challenging but essential. It requires breaking down silos between engineering domains and technical/legal

processes. Our framework and insights lay the groundwork for more resilient autonomous infrastructure, enabling stakeholders to respond effectively and deter future attacks, safeguarding trust and safety in increasingly connected digital and physical systems.

Ethics declaration: Ethical clearance for the research referred to in this paper was not required.

AI declaration: Generative AI in the form of assistive tools within Google Docs was primarily used to reorganize this manuscript. Originally, it was more than twice the length of this current draft. Grammarly was used in the assistance of smoothing writing styles and addressing grammar.

References

Abdo, J.B. and Hossain, L., 2025. Interdisciplinarity of Cybersecurity: Towards Convergence Science. Authorea Preprints.

Almusayli, A., Zia, T. and Qazi, E.-U.-H. (2024) 'Drone Forensics: An Innovative Approach to the Forensic Investigation of Drone Accidents Based on Digital Twin Technology', Technologies, 12(1), p. 11. doi: 10.3390/technologies12010011.

Barnett, M., Samori, I., Griffin, B., Palmer, X.L. and Potter, L., 2023, June. A Commentary and Exploration of Maritime Applications of Biosecurity and Cybersecurity Intersections. In ECCWS 2023 22nd European Conference on Cyber Warfare and Security (No. 1). Academic Conferences and publishing limited.

Casey, E., Barnum, S., Griffith, R., Snyder, J. and van Beek, H. (2015) 'Advancing coordinated cyber-investigations and tool interoperability using a community-developed specification language', Digital Investigation, 12, pp. S102–S110. doi: 10.1016/j.diin.2015.01.013.

Haber, M.J. and Rolls, D., 2024. Indicators of Compromise. In Identity Attack Vectors: Strategically Designing and Implementing Identity Security, Second Edition (pp. 87-107). Berkeley, CA: Apress.

Hartmann, K. and Giles, K. (2016) 'UAV Exploitation: A New Domain for Cyber Power', in Proceedings of the 8th International Conference on Cyber Conflict (CyCon 2016). Tallinn, Estonia: NATO CCD COE Publications, pp. 205–221. Haber, M.J. and Rolls, D., 2019. Indicators of compromise. In *Identity Attack Vectors: Implementing an Effective Identity and Access Management Solution* (pp. 103-105). Berkeley, CA: Apress.

Hwang, S. and Kim, T.S., 2023. An exploratory study on artifacts for cyber attack attribution considering false flag: using Delphi and AHP methods. IEEE Access, 11, pp.74533-74544.

Ibadah, N., Benavente-Peces, C. and Pahl, M.-O. (2024) 'Securing the Future of Railway Systems: A Comprehensive Cybersecurity Strategy for Critical On-Board and Track-Side Infrastructure', Sensors, 24(24), p. 8218. doi: 10.3390/s24248218.

Karnouskos, S., 2011, November. Stuxnet worm impact on industrial cyber-physical system security. In IECON 2011-37th Annual Conference of the IEEE Industrial Electronics Society (pp. 4490-4494). IEEE.

Kim, B., Lee, S., Park, J. and Kim, H. (2021) 'Cybersecurity for unmanned aerial vehicles', IEEE Communications Magazine, 59(8), pp. 34–40. doi: 10.1109/MCOM.001.2000919.

Lian, Z., Shi, P. and Chen, M., 2024. A Survey on Cyber-Attacks for Cyber-Physical Systems: Modeling, Defense and Design. IEEE Internet of Things Journal.

Longo, A., Nicoletti, M., Milazzo, G. and Mendola, S. (2025) 'Cyber–Physical Resilience: Evolution of Concept, Indicators, and Legal Frameworks', Electronics, 14(8), p. 1684. doi: 10.3390/electronics14081684.

Minnaar, A., 2022. Border security: An essential but effective tool in combatting cross-border crime. In The Handbook of Security (pp. 357-378). Cham: Springer International Publishing.

Potter, L. and Palmer, X.L., 2023. Mission-aware differences in cyberbiosecurity and biocybersecurity policies: Prevention, detection, and elimination. In Cyberbiosecurity: A new field to deal with emerging threats (pp. 37-69). Cham: Springer International Publishing.

Pyke, D. (2025) 'Future Rail Security: How DAS and Drones Redefine Infrastructure Protection', Sensonic Blog, 22 July. Available at: https://www.sensonic.com/en/start/future-rail-security-how-das-and-drones-redefine-infrastructure-protection--3286/ (Accessed: 8 September 2025).

Qiu, S., Shao, Z., Wang, J., Xu, S. and Fei, J. (2024) 'Research on Power Cyber-Physical Cross-Domain Attack Paths Based on Graph Knowledge', Applied Sciences, 14(14), p. 6189. doi: 10.3390/app14146189.

Rid, T. and Buchanan, B. (2015) 'Attributing cyber attacks', Journal of Strategic Studies, 38(1–2), pp. 4–37. doi: 10.1080/01402390.2014.977382.

Semertzis, I., Goyel, H. and Stefanov, A. (2024) 'Towards Real-Time Distinction of Power System Faults and Cyber Attacks on Digital Substations Using Cyber-Physical Event Correlation', in Proceedings of the 12th International Workshop on Modeling and Simulation of Cyber-Physical Energy Systems (MSCPES 2024). Los Alamitos, CA: IEEE.

Simons, G., Danyk, Y. and Maliarchuk, T., 2020. Hybrid war and cyber-attacks: creating legal and operational dilemmas. Global Change, Peace & Security, 32(3), pp.337-342.

Soderi, S., Masti, D. and Lun, Y.Z. (2023) 'Railway cyber-security in the era of interconnected systems: A survey', IEEE Transactions on Intelligent Transportation Systems, 24(7), pp. 6764–6779. doi: 10.1109/TITS.2023.3254442.

Strom, B.E., Applebaum, A., Miller, D.P., Nickels, K.C., Pennington, A.G. and Thomas, C.B., 2018. Mitre att&ck: Design and philosophy. In Technical report. The MITRE Corporation.

Weber, L., 2025. ICAO in the 21st century. In The Elgar Companion to the Law and Practice of the International Civil Aviation Organization (pp. 677-692). Edward Elgar Publishing.

Yadav, A., Diedrich, H. and Chowdhury, O. (2023) 'Secure and Trustworthy Sensor Data Fusion for Cyber-Physical Systems', IEEE Internet of Things Journal, 10(15), pp. 12544–12557. doi: 10.1109/JIOT.2023.3244775.

Yi, C.G. and Kim, Y.G., 2024. Hypothesis Generation Model for Cyber Threat Hunting. IEEE Communications Magazine, 62(10), pp.110-116.

Zero Trust Architecture for UAV/Rail Logistics Ecosystem

Isaac Ojeh[1], Xavier Palmer[2] and Lucas Potter[2]
[1]MorphHats InfoSecure, Waterloo, Canada
[2]BiosView Labs, Dayton, Ohio

morpheus@morphhats.com
Biosview1@proton.me

Abstract: As unmanned aerial vehicles (UAVs) and smart rail systems become increasingly integrated into logistics and supply chain operations, ensuring secure, resilient communication and coordination between these components is critical. Traditional perimeter-based security models are no longer sufficient for such dynamic, distributed ecosystems. This research proposes a Zero Trust Architecture (ZTA) tailored to the UAV/rail logistics domain, addressing the challenges of securing multimodal transport nodes, autonomous coordination, and real-time data exchange across untrusted and heterogeneous networks. Unlike conventional models that assume implicit trust within a network boundary, ZTA continuously validates every user, device, and service attempting to interact with the system, enforcing strict access controls and real-time behavioral monitoring. This paper adopts an exploratory approach to assess the feasibility and impact of applying Zero Trust principles to a complex, multimodal logistics ecosystem. Rather than presenting a finalized solution, the work investigates key design considerations, architectural trade-offs, and implementation challenges unique to UAV and rail system integration. The proposed architecture is designed to authenticate and authorize interactions across UAVs, rail infrastructure, edge nodes, and command-and-control centers using principles of least privilege, micro-segmentation, and continuous verification. Machine learning models are incorporated to detect anomalies in system behavior, such as route deviations or unauthorized data flows, enabling rapid threat identification and mitigation. Furthermore, this approach integrates policy-based access control with identity-aware proxies to enforce trust decisions dynamically and contextually based on device posture, geolocation, and real-time mission parameters. The implications of this work extend beyond protecting individual assets; they enable the secure orchestration of complex, time-sensitive logistics chains involving autonomous and semi-autonomous agents. This work demonstrates that implementing Zero Trust in UAV/rail logistics is both technically viable and operationally beneficial, offering a path toward resilient, adaptive infrastructure capable of withstanding evolving cyber threats without compromising performance or interoperability.

Keywords: Zero trust architecture, UAV logistics, Rail infrastructure security, Autonomous systems, Secure communication

1. Introduction

Modern supply chains are increasingly integrating unmanned aerial vehicles (UAVs), drones with smart rail systems to boost efficiency (Debevec, 2019; Kafetzis et al, 2020; Park et al, 2022; Chen et al, 2025; Jahani et al, 2024). Drones are now critical tools in delivery, surveillance, and inspection across the Logistics and Supply Chain (LSC) sector (Jahani et al. 2024). Simultaneously, U.S. rail networks are becoming "smart," using advanced technologies to optimize operations. This convergence envisions scenarios like drones transferring high-priority packages to or from trains, aerial rail line inspection, and real-time data exchange for synchronized operations. This multimodal integration promises significant gains but also expands cyber threat attack surfaces. Securing transportation infrastructure is paramount. U.S. freight railroads move over billion tons of freight annually, including energy rich, polluting, and hazardous materials, making them attractive cyberattack targets (Kour et al, 2023; Ho et al, 2024). Positive Train Control (PTC) implementation by 2020 connected locomotives, wayside devices, and control centers into an IoT-like network, improving safety but introducing vulnerabilities (Swain 2022). Cyberattacks on train control systems could be catastrophic and mass casualties are difficult to rule out. Incident discussions at the intersection of cybersecurity and railways appear largely conceptual, but enough considerable attacks have occurred in numerous countries since at least 2008, affecting both passenger travel and industry, to press the question of how they may broaden in effect (Kour et al, 2022; Thron et al, 2024; Fernandes et al, 2025).

Broadcast-controlled UAVs also face threats like GPS spoofing, jamming, and control link hijacking. Protecting drone data is a growing concern (McNabb, 2025). The challenge is securing this distributed, heterogeneous drone and rail ecosystem across untrusted networks. Traditional perimeter-based security models are unsuited for this scenario. They assume internal network entities are trustworthy, focusing on external threats. However, in a combined UAV-rail logistics environment, a single protected perimeter vanishes; drones use wireless links, rail systems interface with cloud services, and data traverses public networks. Insider threats or compromised nodes are also risks. The industry is adopting a Zero Trust (ZT) paradigm: “Never trust, always verify” (McNabb 2025). Every interaction is treated as potentially hostile; all users, devices, and applications must continuously prove identity, integrity, and authorization (Rose et al. 2020). Zero Trust Architectures (ZTAs) aim to limit breach impact and prevent lateral movement by enforcing least privilege access and constant verification, assuming

attackers may already be present (Rose et al. 2020). This paper explores applying ZTA principles to integrated UAV and rail logistics conceptually, drawing from literature. We identify key security requirements, design considerations, and implementation challenges through an exploratory approach grounded in existing research and industry practices. We aim to outline a plausible architecture and its components and discuss how it could enhance UAV/rail operations. We draw on knowledge from industrial control system (ICS) security, IoT/edge computing, and autonomous systems, highlighting areas for further research.

The paper is organized as follows: Section 2 reviews Zero Trust principles and cybersecurity in UAVs and rail systems. Section 3 looks at Gaps and Motivation for a Zero Trust Approach. Section 4 Proposed Zero Trust Architecture for UAV–Rail Logistics. Section 5 Design Considerations and Challenges. Section 6 discussed all of the above before leading into a conclusion. Overall this paper helps to bring attention to this area of intersection and stimulate exploration.

2. Background and Related Work

2.1 Zero Trust Security Principles

ZT is a shift from traditional network security, where no user or device is trusted, even inside the network perimeter; every access request must be verified (Kindervag 2010; Rose et al. 2020). This philosophy emerged over a decade ago due to sophisticated threats bypassing perimeter defenses and has since gained widespread acceptance, including government endorsement (NIST SP 800-207 in 2020, U.S. DoD Zero Trust Strategy in 2022) (Rose et al, 2020: Choi, 2025). Its core tenets are (Rose et al. 2020):

- **Assume Breach:** Defenses focus on internal monitoring, assuming malicious actors may already be present.
- **Least Privilege Access:** Entities receive minimum necessary access, minimizing impact if compromised.
- **Continuous Verification:** Every access is authenticated and authorized in real-time, with no persistent trust. Many ZT aspects (e.g., multi-factor authentication) exist in modern defenses, but ZT formalizes this for every interaction (Benestelli and Kambic 2022).
- **Micro-Segmentation:** Systems are divided into small trust zones, preventing compromise in one segment from affecting others. This can be achieved through software-defined networking or host-based firewalls.
- **Assume External Networks:** Internal traffic is treated as hostile, requiring practices like end-to-end encryption and rigorous integrity checking.

NIST defines a ZTA using a Policy Engine, Administrator, and Enforcement Points. Dynamic, risk-based authorization, fed by identity management, authentication, and continuous monitoring, adapts access based on context like device health and location. Implementing ZTA in OT/ICS, such as railway signaling, is challenging due to legacy platforms, real-time needs, and protocols lacking security (Benestelli & Kambic 2022). Continuous authentication or encryption could disrupt operations. Shared accounts or no authentication on OT devices also conflict with ZTA. Therefore, a direct "lift-and-shift" of IT Zero Trust is impractical for rail or UAV systems; instead, gradual segmentation, encryption overlays, and jump hosts are suggested (Benestelli & Kambic 2022).

2.2 Cybersecurity in UAV and Rail Systems

UAVs, or drones, are widely used across various sectors, facing increasing physical and cyber threats, especially in sensitive applications. Their reliance on communication links for command and control makes them vulnerable to eavesdropping, jamming, and spoofing. Documented instances, like GPS spoofing and command-link hijacking, highlight these threats, which can misdirect drones or allow attackers to seize control. The rise in unauthorized drone takeovers and their use in cyber-physical attacks underscores the critical need for robust UAV security measures, as emphasized by Haque et al. (2024), to prevent espionage, data exfiltration, kinetic attacks, and infrastructure disruption.

2.3 Foundational Security Principles and Emerging Solutions

At its core, one can glean that securing UAVs necessitates ensuring the authenticity and integrity of all commands and data transmitted between the drone and its ground station (Michailidis et al, 2022; Haque et al, 2024). Without these foundational principles, any communication is susceptible to manipulation or impersonation. Researchers and industry experts have proposed and are actively developing a range of solutions to address these challenges. Mutual authentication between the drone and its ground station is a cornerstone

of secure UAV operations. This process verifies the identity of both parties before any communication or command execution, preventing unauthorized entities from interacting with the UAV. Encryption of control signals is another vital layer of defense, scrambling the data so that even if intercepted, it remains unreadable and unusable to unauthorized parties. Furthermore, the deployment of sophisticated IDS specifically tailored for UAV networks is crucial. These systems monitor network traffic and behavior for anomalies that could indicate a cyberattack, enabling timely detection and response (Valikhanli, 2024; Islam et al, 2025).

2.4 The Rise of Zero Trust in UAV Networks

The Zero Trust security model is transforming UAV security by shifting from perimeter-based trust to continuous verification of every action, regardless of its origin. It is marked by value in critiquing every action to pre-empt misuse and leaving behind default trust, even for familiar components. SpiderOak's Zero Trust platform exemplifies this, ensuring that applications authenticate all outbound drone messages, thus mitigating spoofing and hijacking (McNabb, 2025). This model becomes critical as regulations like the FAA's BVLOS operations rules open new airspace for complex drone applications. BVLOS operations necessitate automated data sharing, expanding the attack surface. In this dynamic environment, Zero Trust's continuous verification of interactions will be indispensable for the future of UAV operations, which hinges on robust and adaptive security frameworks.

2.5 Railway Cybersecurity

Historically prioritizing safety and reliability, the rail industry is now confronting cybersecurity challenges as systems become hyper-connected (Soderi et al. 2023). Modern railways incorporate diverse networked components, including computer-controlled signaling systems, wireless communication for train control (e.g., PTC in the U.S., ETCS in Europe), and SCADA systems for power and infrastructure management. While these advancements boost efficiency and safety, preventing collisions and optimizing routes, they also expose rail operations to cyberattacks that could disrupt service or cause accidents.

The U.S. rail system's PTC highlights its vulnerability, as interconnected components like locomotives and trackside devices are susceptible to cyberattacks that could inject false data or malicious commands, potentially causing derailments or collisions. This is especially critical given freight rail's transport of hazardous materials, where a cyber incident could have severe environmental and public health consequences. Real-world events, such as the 2021 cyberattack on Iran’s rail system and the 2022 targeting of Belarusian railways by hacktivists, underscore these risks. In response, agencies like the U.S. Transportation Security Administration (TSA) have mandated cybersecurity enhancements for major rail operators, aligning with Zero Trust principles. Soderi et al. (2023) emphasize the need for a balance between security and reliability in railway cybersecurity, highlighting the cyber attack surface and the importance of frameworks like IEC 62443 and CENELEC TS 50701. The interdependence of safety and security in rail operations means a cyber compromise can directly cause physical harm, necessitating robust authentication, authorization, and continuous monitoring for anomalies in train control systems.

3. Gaps and Motivation for a Zero Trust Approach

UAV and rail systems share security vulnerabilities due to increased connectivity and autonomy, which traditional methods struggle to address. ZTA offers a solution for these distributed, dynamic networks, with some ZTA applications explored in each domain separately (Haque et al. 2024; Benestelli & Kambic 2022). However, little work has focused on the convergence of UAV and rail logistics, where a holistic security approach is critical. Consider a hypothetical UAV-rail operation: an autonomous drone exchanges parts or data with a moving freight train. Both must authenticate with each other and a central control system. These transactions occur over various, potentially insecure networks (local wireless, cellular/satellite, dedicated rail telecom). This complex, cross-organizational scenario highlights ZTA's value. By authenticating every interaction, enforcing encryption, and continuous monitoring, ZTA can prevent a compromised component from disrupting the logistics chain. This research aims to outline a ZT-based model for UAV/rail logistics, identifying key considerations and challenges. The next section proposes conceptual architecture integrating elements like identity management and anomaly detection.

4. Proposed Zero Trust Architecture for UAV–Rail Logistics

This architecture applies Zero Trust principles to an integrated UAV-rail logistics system, as shown in **Figures 1, 2 and 3**. A centralized Policy Engine/Authority makes all access decisions. Every device or service, from UAVs to operator workstations, must authenticate and obtain authorization from this engine before critical interaction.

Policy Enforcement Points (PEPs) are deployed strategically on devices, message brokers, and network gateways. These PEPs mediate requests, consulting the Policy Engine, which uses pre-defined security policies and real-time context to grant or deny actions.

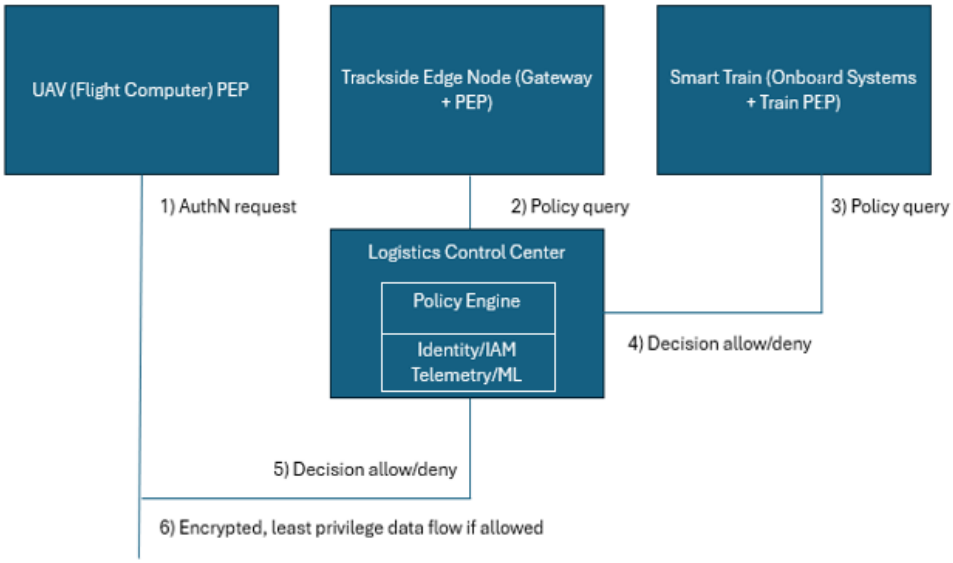


Figure 1: This presents a simple schematic illustrating a conceptual Zero Trust Architecture for a UAV-Rail Logistics ecosystem. In this architecture, a Policy Enforcement Point (PEP) is crucial for managing access. All data paths within the system are encrypted, and every access request undergoes evaluation against established policies, adhering to principles of least privilege and continuous verification, with robust authentication (AuthN) being a fundamental requirement

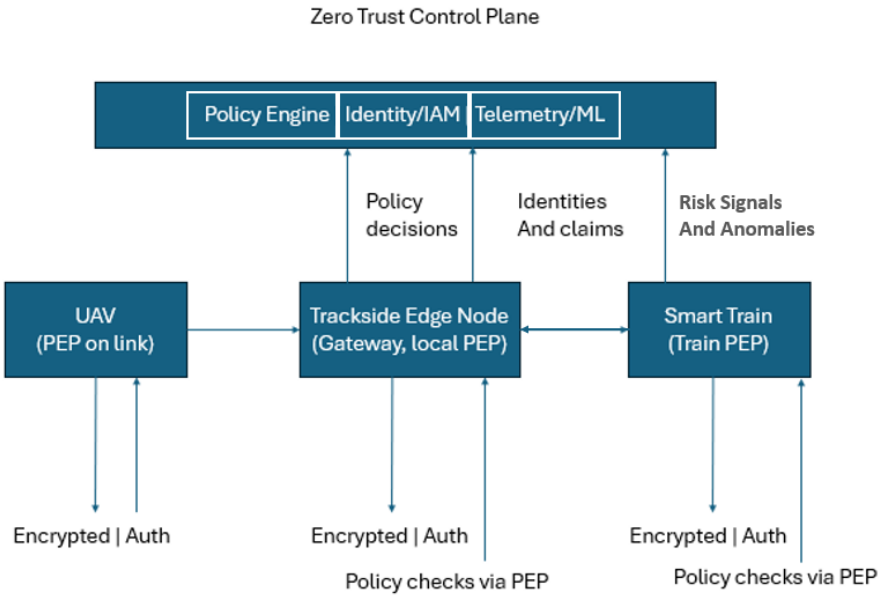


Figure 2: This presents an infographic-style overview of the Zero Trust Architecture for a UAV/Rail Logistics Ecosystem. This architecture defines distinct zones and micro-segments, including a UAV zone for airborne links with constrained, rotating tokens, an Edge zone featuring a gateway Policy Enforcement Point (PEP), local cache, and short-lived sessions, and a Train zone for onboard systems with segmented buses and strict PEP. A central Control plane manages policy, identity, and analytics. Key Zero Trust principles are applied, such as "never trust, always verify" where every request routes through a PEP to the Policy Engine, and "least privilege" with tokens scoped to

single tasks and short durations. Micro-segmentation ensures isolation between UAV, Edge, and Train components to prevent lateral movement, while continuous verification considers posture, location, time window, and mission context to drive access decisions

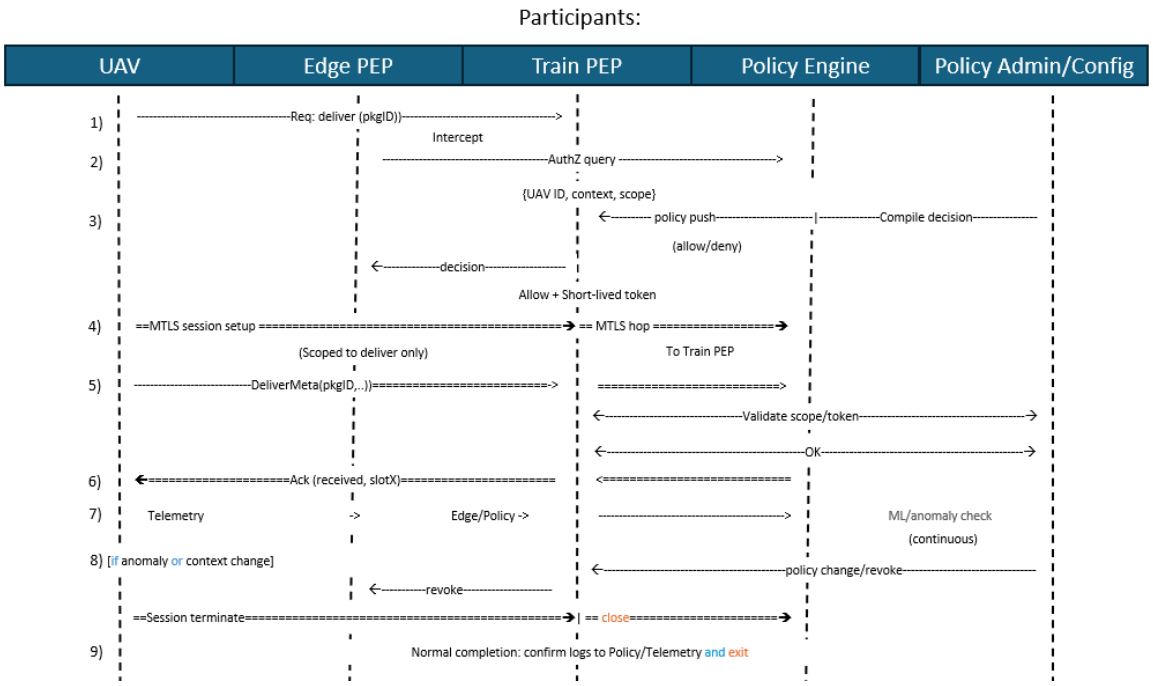


Figure 3: UAV-to-Train Delivery Sequence Diagram. All steps are authenticated, authorized, and continuously verified. Tokens are least-privilege and time-bound. The diagram serves as a high-level architectural sequence rather than a fully formal UML sequence diagram. The direction of messages is indicated both textually and by the orientation of arrows, and activation bars have intentionally been omitted to enhance readability

Figures 1-3 depict the Zero Trust Architecture for UAV-integrated smart rail logistics, asserting that all entities (drone, train, node, user) are untrusted until verified (Haque et al. 2024; IoT Security Institute 2025). This aligns with least privilege, micro-segmentation, continuous authentication, and policy-driven access, where all communications require central authentication and authorization. Secure channels and continuous identity verification are used by UAVs, smart trains, trackside edge nodes, and the logistics control center. Even within the "Railway Infrastructure Zone," explicit authentication and least privilege limit compromise.

The Zero Trust Architecture (ZTA) for autonomous logistics centers on five key areas. First, Device Identity and Authentication ensures each UAV and rail node possesses a unique cryptographic identity and performs continuous, mutual authentication. Second, Micro-Segmentation and Network Security divides the network into tightly controlled micro-perimeters, with each segment having its own Policy Enforcement Point and allowing communication only between verified identities. Third, the Policy Engine and Contextual Access Control acts as a central decision point, applying dynamic, least-privilege policies based on real-time context and enterprise rules, enforced via identity-aware proxies. Fourth, Continuous Monitoring and Anomaly Detection utilizes machine learning to analyze telemetry, flagging anomalies in real-time for mitigating action and employing Explainable AI to understand model outputs. Finally, Secure Communication and Data Integrity mandates end-to-end encryption and integrity checks for all data exchanges, with critical commands cryptographically signed to prevent tampering. Additionally, Resilience and Fail-Safes are crucial, incorporating redundant Policy Engine instances and defining fail-safe behaviors to prevent ZTA controls from becoming single points of failure.

The proposed architecture establishes a web of trust for UAV and rail components. It addresses the network's heterogeneous and shifting nature by using identity as a constant anchor, rather than IP addresses or physical network segments. Each major subsystem continuously asserts its identity and authorized actions. The next section discusses critical design considerations and challenges for such a viable system.

5. Design Considerations and Challenges

Implementing Zero Trust (ZT) in combined UAV and smart rail environments faces challenges from heterogeneous devices, legacy systems, and connectivity. Newer rail equipment and UAVs contrast with older rail signal controllers (Haque et al. 2024). Solutions like edge proxies for low-power devices introduce trust dependencies (Benestelli & Kambic 2022). The rail industry's long technology cycles complicate integrating older protocols. Bastion hosts can secure legacy components, but 24/7 operations hinder upgrades. Intermittent UAV and train connectivity challenge ZT's continuous verification. Offline tokens can be security-sensitive. Latency for critical train decisions can be mitigated by hierarchical Policy Engines or edge nodes. The ZT infrastructure itself is a high-value target, requiring robust hardening and redundancy for identity systems and policy engines, plus crucial monitoring. Defining policies for complex UAV-rail systems is difficult, needing clear taxonomy, simulation, and human oversight. Collaboration between railway and drone security teams is vital. ML-based anomaly detection improves security but risks false positives, which can disrupt safety-critical operations. Tuning with domain knowledge and multi-tiered responses are necessary. Adversaries might try to evade ML models, so diverse detection and secure data pipelines are important. Compliance with regulations (e.g., NIST, TSA) is critical; ZT principles align, but demonstrating compliance in a dynamic ZTA requires strong logging and reporting. Interoperability across organizations is vital, requiring federated trust mechanisms like SAML/OAuth or PKI. Securing UAV-rail logistics with ZT demands careful attention to device constraints, legacy systems, and operational complexities. Incremental, well-tested deployments are recommended, mirroring other OT domains. The next section explores benefits and future research foci.

6. Discussion

Zero Trust Architecture (ZTA) offers a path to enhance cybersecurity in integrated UAV-rail operations, shifting from reactive to proactive security and significantly improving resilience against advanced threats by containing breaches. Specifically, ZTA limits lateral movement, unlike traditional networks, where a compromised drone could infiltrate the rail network, applies least privilege, and blocks unusual actions. Continuous monitoring by operators and properly configured tools also quarantines anomalous devices, thereby creating a "secure by design" system ideal for critical infrastructure facing nation-state threats. Beyond internal security, ZTA can excel at facilitating secure interoperability between organizations such as rail operators and drone delivery providers, and it establishes a common security framework with shared identity and policy, avoiding full network integration and increasing vulnerability. This selective sharing, aligned with cloud computing Zero Trust approaches, supports new business models like a drone delivering to a smart locker, with both authenticating via a common ZT service. Such federated ZT arrangements are expected to grow as industries converge. Admittedly, while new security measures can raise operational concerns, the goal of this architecture is minimal intrusion. Fast authentication and automated policy decisions ensure smooth operations until a security issue arises. However, strict access control may occasionally block legitimate actions, necessitating strong communication between security and operations teams and continuous policy review. Despite this, ZTA can maintain or improve operational continuity by preventing incidents and avoiding costly delays (Rose et al. 2020). Furthermore, Zero Trust's adaptability and future-proofing stem from its focus on identity, authentication, and minimal access, making it agnostic to specific threats. Incorporating machine learning for anomaly detection and context-aware policies further enhances its ability to adapt to evolving threats. While not a "silver bullet" against all attacks, ZTA makes it harder for attackers to cause damage and provides robust defense tools. Nevertheless, limitations and open questions include the human factor, where frequent re-authentication could lead to alert fatigue. Cost and complexity are also concerns, especially for smaller operators, suggesting a need for phased or managed-service approaches (McNabb, 2025). Rigorous testing and validation, through additional and varied cyber-physical simulations (Soderi et al. 2023), are crucial to assess ZTA's performance under attack. In conclusion, Zero Trust offers a promising solution for securing complex logistics ecosystems with UAVs and smart rail systems. Its success depends on cross-sector collaboration among cybersecurity, rail engineering, and autonomous systems experts to refine the architecture and ensure it enhances safety and efficiency. This collaborative approach is vital because, given the growing and intersecting biocrime avenues and the widening attack surfaces in the bioeconomy, how we treat our technological intersections is increasingly important; otherwise, our preparedness remains in question (Elgabry, 2022; Potter and Palmer, 2023; Elgabry and Johnson, 2024). Additionally, it is important that the human element behind our training and operation of technological systems and components not go underexamined (Nurse et al., 2018; Dupont and Holt, 2022; Potter and Palmer, 2021). Ultimately, an open mind towards our combined human and technological vulnerabilities as we reexamine our systems of trust will lead to better planning, especially as greater reliance on rail and UAV systems continues to be adopted for moving materials that are central to our ways of life.

7. Conclusion

Integrating UAVs with smart rail logistics faces cyber threats from both domains. Our exploratory study proposes a Zero Trust Architecture (ZTA) as a unifying security strategy. Based on "never trust, always verify," and current cybersecurity, we outlined a conceptual ZTA for continuous authentication and authorization among drones, trains, trackside devices, and control systems. ZTA principles like least privilege, micro-segmentation, and continuous monitoring can mitigate risks, from spoofed drones to compromised signaling, reducing cyberattack likelihood and preventing physical harm or disruptions. Our conceptual architecture presents challenges like legacy system compatibility, network latency, and robust anomaly detection. While powerful, ZTA implementation requires careful engineering, iterative testing, and a nuanced approach respecting real-world, safety-critical operations. These challenges are not insurmountable but highlight the need for focused development, such as lightweight cryptography for low-power devices or improved ML for anomaly detection. Zero Trust is not merely theoretical; elements are emerging, with stakeholders recognizing its value. Companies rolling out ZT platforms for drones and governments pushing critical infrastructure operators towards ZT-like measures, reinforce this exploration's timeliness. That said, our work uniquely addresses the security intersection of UAVs and railways. As a conceptual and exploratory paper, it provides a foundation for future research, including prototyping parts of architecture discussed. In summary, a ZTA for UAV and rail logistics is feasible and advantageous, offering resilient, adaptive infrastructure against evolving cyber threats without sacrificing interoperability or performance. As autonomous agents become prevalent, security architectures must evolve. Zero Trust provides a compelling blueprint, enabling the benefits of integration and autonomy in logistics while keeping risks acceptable. We encourage further investigation, testing, and refinement to move closer to secure, zero-trust logistics chains.

Ethics declaration: Ethical clearance for the research referred to in this paper was not required.

AI declaration: Generative AI in the form of assistive tools within Google Docs was primarily used to condense this manuscript and summarize sections. Originally, it was more than twice the length of this current draft. Grammarly was used in the assistance of smoothing writing styles and addressing grammar.

References

Benestelli, B. & Kambic, D. (2022). IT, OT, and ZT: Implementing Zero Trust in Industrial Control Systems. Carnegie Mellon Software Engineering Institute Blog, 18 July 2022. Available: https://www.sei.cmu.edu/blog/it-ot-and-zt-implementing-zero-trust-in-industrial-control-systems/.

Chen, M., Wang, H., Zhang, X., Li, Y., Ma, X. and Fan, Y., 2025. The Application and Future Prospects of Drone Vision Technology in the Railway Industry. Journal of Technology Innovation and Engineering, 1(1), pp.39-44.

Choi, Y.B., 2025. A Certainty-Based Approach to Implementing Zero Trust Architecture Using NIST SP 800-207 and NIST SP 1800-35. *KAUPA Letters*, *13*(1), p.11.

Debevec, R., 2019. A smart UAV platform for railroad inspection.

Dupont, B. and Holt, T., 2022. The human factor of cybercrime. *Social science computer review*, *40*(4), pp.860-864.

Elgabry, M., 2022. *The future of biotechnology crime: are we prepared for it?* (Doctoral dissertation, UCL (University College London)).

Elgabry, M. and Johnson, S., 2024. Cyber-biological convergence: a systematic review and future outlook. Frontiers in Bioengineering and Biotechnology, 12, p.1456354.

Fernandes, T., Magalhães, J.P. and Alves, W., 2025. Cybersecurity in Smart Railways: exploring risks, vulnerabilities and mitigation in the data communication services. Green Energy and Intelligent Transportation, p.100305.

Haque, E., Hasan, K., Ahmed, I., Alam, M.S., & Islam, T. (2024). Enhancing UAV Security Through Zero Trust Architecture: An Advanced Deep Learning and Explainable AI Analysis. arXiv preprint arXiv:2403.17093.

Ho, A., Giannopoulos, D., Pilorgé, H. and Psarras, P., 2024. Opportunities for rail in the transport of carbon dioxide in the United States. Frontiers in Energy Research, 11, p.1343085.

IoT Security Institute (2023) Implementing Zero Trust Architecture for Resilient Enterprise Security. (Online) Available at: https://iotsecurityinstitute.com/iotsec/iot-security-institute-cyber-security-articles/200-implementing-zero-trust-architecture-for-resilient-enterprise-security [Accessed 1 September 2025].

Islam, M.D., Mahmoud, A.S. and Sheltami, T.R., 2025. AI-Enhanced Intrusion Detection for UAV Systems: A Taxonomy and Comparative Review. *Drones (2504-446X)*, *9*(10).

Kindervag, J., 2010. Build security into your network's dna: The zero trust network architecture. *Forrester Research Inc*, *27*, pp.1-16.

Kour, R., Patwardhan, A., Thaduri, A. and Karim, R., 2023. A review on cybersecurity in railways. Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit, 237(1), pp.3-20.

Jahani, H., Khosravi, Y., Kargar, B., Ong, K.-L., & Arisian, S. (2024). Exploring the role of drones and UAVs in logistics and supply chain management: a novel text-based literature review. International Journal of Production Research. DOI: 10.1080/00207543.2024.2373425

Kafetzis, D., Fourfouris, I., Argyropoulos, S. and Koutsopoulos, I., 2020, September. UAV-assisted aerial survey of railways using deep learning. In 2020 International Conference on Unmanned Aircraft Systems (ICUAS) (pp. 1491-1500). IEEE.

Kindervag, J. (2010). No More Chewy Centers: The Zero Trust Model of Information Security. Forrester Research Inc., Cambridge, MA.

Michailidis, E.T., Maliatsos, K., Skoutas, D.N., Vouyioukas, D. and Skianis, C., 2022. Secure UAV-aided mobile edge computing for IoT: A review. IEEE Access, 10, pp.86353-86383.

McNabb, M. (2025) Zero trust takes flight: SpiderOak's cybersecurity platform shields drone data from spoofing and theft. DroneLife, 8 April. Available at: https://dronelife.com/2025/04/08/zero-trust-takes-flight-spideroaks-cybersecurity-platform-shields-drone-data-from-spoofing-and-theft/

Nurse, J.R., 2018. Cybercrime and you: How criminals attack and the human factors that they seek to exploit. *arXiv preprint arXiv:1811.06624*.

Park, Y.M., Tun, Y.K., Han, Z. and Hong, C.S., 2022. Trajectory optimization and phase-shift design in IRS-assisted UAV network for smart railway. IEEE Transactions on Vehicular Technology, 71(10), pp.11317-11321.

Potter, L. and Palmer, X.L., 2021, April. Human factors in biocybersecurity wargames. In *Future of information and communication conference* (pp. 666-673). Cham: Springer International Publishing.

Potter, L. and Palmer, X.L., 2023. Mission-aware differences in cyberbiosecurity and biocybersecurity policies: Prevention, detection, and elimination. In *Cyberbiosecurity: A new field to deal with emerging threats* (pp. 37-69). Cham: Springer International Publishing.

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero Trust Architecture (NIST Special Publication 800-207). National Institute of Standards and Technology, U.S. Department of Commerce, Aug. 2020. DOI: 10.6028/NIST.SP.800-207

Soderi, S., Masti, D., & Zacchia Lun, Y. (2023). Railway Cyber-Security in the Era of Interconnected Systems: A Survey. IEEE Transactions on Intelligent Transportation Systems, 24(7), 6764-6779. DOI: 10.1109/TITS.2023.3254442

Swain, C. (2022). The Emerging Cyber Threat to the American Rail Industry. Lawfare (Online), 20 Oct 2022. Available: https://www.lawfaremedia.org/article/emerging-cyber-threat-american-rail-industry.

Thron, E., Faily, S., Dogan, H. and Freer, M., 2024. Human factors and cyber-security risks on the railway–the critical role played by signalling operations. Information & Computer Security, 32(2), pp.236-263.

Valikhanli, O., 2024. UAV networks DoS attacks detection using artificial intelligence based on weighted machine learning. *Results in Control and Optimization*, *16*, p.100457.

Policing Vulnerability to Criminal *Persuasive Technology* Use on Futurist Moon Bases

Tim Pappa
Capitol Technology University, Laurel, Maryland, USA

tpappa@captechu.edu

Abstract: There have been increasing demonstrations of criminal enterprises normalizing the adaptation of *persuasive technologies* to target and victimize people all over the world online and offline. Persuasive technologies include any software or hardware that we interact with that can influence our behaviors, including changing our behaviors or maintaining our behaviors, making people vulnerable to behavioral or technical exploitation of our information and communication. This vulnerability may be heightened in isolated environments like futurist Moon Bases, suggesting there may be a unique victimology for people living and working on a Moon Base. To counter this activity, in theory, policing criminal persuasive technology use on a Moon Base environment must consider the behavioral design of persuasive technologies and this unique victimology. This practitioner's paper will discuss the shifting research direction of persuasive technologies toward deception, since the introduction of this model focused on improving people's health almost two decades ago, suggesting that generative artificial intelligence and machine learning will accelerate adaptive, deceptive persuasive technologies. This paper will visualize theoretical scenarios describing how criminal enterprises could victimize people on Moon Bases with persuasive technology design specific to that environment. This paper provides a comparative example when considering Arctic contexts as similar isolated environments. These futurist theoretical scenarios provide some demonstrations of criminal behaviors using persuasive technologies to victimize people, highlighting the unique victimology of users in isolated environments of interest to criminals and nations alike.

Keywords: Persuasive technology, Crime, Moon bases, police, Deception, Artificial intelligence

1. Introduction

This theoretical practitioner's paper suggests that people living and working in isolated environments such as futurist Moon Bases may be more behaviorally vulnerable to exploitation than anywhere on Earth, complicating the potential extraordinary challenges of law enforcement to police crimes against people.

This paper particularly concentrates on the adaptive use of persuasive technologies to exploit people, especially in isolated environments where people would depend heavily on communication and online infrastructure. Persuasive technologies can be any software or hardware that we interact with online that influences our behaviors, including changing our behaviors and maintaining our behaviors (Oyibo, 2024; Idrees et al., 2024; Pintar and Erjavec, 2021; Ramachandran and Canny, 2008). Examples could include everything from social media to fitness applications like Fitbit to games like Candy Crush and language learning applications like Duolingo.

Hermida (2006) challenged the expectation at the time that isolated conditions in outer space would result in a high rate of "criminal and deviant conflicts". But other than an extraordinary incident where a Russian astronaut physically assaulted male astronauts and sexually assaulted and harassed a female astronaut, astronauts in the International Space Station are continuously monitored. Even prison inmates are monitored less. Hermida suggested this continuous monitoring established after the incident where there were multiple assaults could be considered an extraordinary deterrent. Hermida also highlights that generally the cohorts of astronauts selected for training and sent to space do not reflect the demographics and backgrounds of typical classes of criminals. There are sociological and psychological explanations for how this sample of astronauts is not expected to be as predisposed to criminal activity. The visualization in this paper explores the very likely environment at some point in the future where there is a larger population of people in outer space, where astronauts will be in the minority. A futurist Moon Base could plausibly include hundreds of people from civilian and military organizations.

The Arctic is a helpful example of how an isolated environment will continue to grow with people as infrastructure grows, especially when commercial and civilian organizations begin to establish their roles in those environments. The number of people deploying to or establishing an operating presence in the Arctic region is steadily growing, which has resulted in a growing need for infrastructure to support them. That infrastructure has included Starlink Wi-Fi access not just for facilitating military and scientific activities, but for the morale of personnel deployed to these isolated environments. Starlink provided similar internet infrastructure to personnel in Antarctica for personal use. There is arguably heightened use with this available infrastructure provided by commercial organizations because people are at times isolated or far from home in stressful and contest geopolitical locations. Outer space has already been characterized in a similar way when imagining

futurist Moon Bases. Persuasive technologies can drive our choices and interaction online, potentially making us vulnerable to collection or some other behavioral or technical exploitation of our accounts or information (Tariq, 2024).

An exploratory National Intelligence University study found among a sample of American military personnel who deployed in the past few years to similar "isolated intelligence domains" like the Arctic and Antarctica that younger personnel were more likely to use gaming applications on their smartphones than middle-aged personnel (Rudolph, 2024). While that finding may not be surprising, what was surprising was that middle-aged personnel tended to feel less lonely than younger personnel. There are several possible explanations for this preliminary finding, but this more recent research on the effects of isolation and loneliness on deployed military personnel and their use of persuasive technologies is crucial considering the growing population of military personnel deployed to the Arctic region will likely be younger. These kinds of studies are rare.

Some of the early anthropological and ethnographic research conducted in Arctic communities in the 1970s focused closely on the effects of the environment and other social factors on communities that were isolated and mobile in this region (Savishinsky, 1971). Many of those communities were indigenous. That starting point provided some context for exploring the effect of that same environment on people visiting that region for shorter or longer durations. Social psychologist Veronika Sharok (2020) conducted assessments of the psychological characteristics of the kinds of people who adapt best to the Arctic environment. Sharok found that generally the people who expressed the greatest interest in deploying to the Arctic adapted better to that environment and the people who stayed for a longer duration maintained a better relationship with their families. The people who tended to "fly-in fly-out" which may be the best approximation of military personnel on short-term deployments did not adapt well to the Arctic conditions and generally evaluated themselves to be less happy and less connected with their families at home. Sharok highlighted consistent findings in past research and in her research that found that interpersonal relationships seemed to have the greatest influence on coping behaviors and psychological well-being in such a harsh, isolated environment. Russian researchers more recently highlighted how the development of industries and infrastructure in the Arctic region has not thoroughly considered and measured the "psychophysiological" impact of working conditions and environment on people (Bashkireva et al., 2019). This environment makes "special demands" on people beyond their physical health, especially in terms of their mental health. The researchers noted as an example of an unexplored factor on people in this region the effect of magnetic and solar storms in this region on people living in this region. In some cases, someone's circadian rhythm can change.

Some of the same harsh working conditions and environment described above could be similar on futurist Moon Bases. The author recognizes that there is limited research into the effects of outer space environments on people and groups of people, but the references in the Arctic provide some perspective on how persuasive technologies could be used to criminally exploit people on futurist Moon Bases.

Although this theoretical paper explores how law enforcement may attempt to police the vulnerability of persuasive technology use on a futurist Moon Base, the author will only briefly discuss this context of criminology. There has been a growing subdiscipline of criminology in recent years that has mostly focused on the environmental harms people could create in outer space, such as mining companies in space trying to mine asteroids for minerals or space debris (Burns and Lampkin, 2024; Lampkin and Wyatt, 2023; Lampkin and McClanahan, 2023). This subdiscipline of "astro-green criminology" or space criminology has considered "predictive models of human space communities" based on our understanding of people and their behaviors in similar or artificial environments, such as the Arctic. The ongoing research into human factors in outer space environments has consistently highlighted the "mental processing of such extreme experiences" for people living in those physical and social environments (Pagnini et al., 2023). This theoretical practitioner's paper will introduce this unique vulnerability or perhaps victimology in this imagined outer space environment. The author has a background in law enforcement, researching and analyzing how people interact online and how people can be victimized online when using applications personalized for them.

This paper will be organized as follows. First, the author will discuss more background on the development of persuasive technology as a conceptual framework for understanding how people interact with technologies and how they can be influenced behaviorally. This literature review will explain some of the apparent shift in the design and research of persuasive technologies since this conceptual framework was developed more than two decades ago, to contexts of deception use. Second, the author will visualize two scenarios on a futurist Moon Base where persuasive technology design and application can be used by criminals to victimizers users living and working on a futurist Moon Base. Finally, this paper will discuss how these vulnerabilities could be policed,

including if military and intelligence services were also operationalizing persuasive technology design to target victims.

2. The Shift in Persuasive Technology Design and Research to Deceptive Contexts

Social scientist B.J. Fogg (1998, 2002) first named persuasive technology, which he defined as any interactive computing system designed to change or maintain someone's attitudes or behaviors. Fogg coined the phrase "captology" to characterize the limited space at the time in the early 2000s where technologies and persuasion overlapped in peoples' lives, anywhere there was human-computer interaction. Fogg emphasized that captology focuses on the *planned* persuasive effects of the technology, not on the 'side effects' of that technology.

While any technology is going to influence a user in some way, for a technology to be called "persuasive" that persuasion or influence must be intentional (Hamari, Koivisto, and Pakkanen, 2014). Technology is persuasive because it is designed to guide someone toward an attitude or behavior change or toward maintaining an attitude or behavior change.

Fogg (2009) recognized the social environment online and offline also shaped how responsive a user with complex behaviors might be to suggestions built into persuasive technology, but he wrote that he wanted to know how motivated someone was to perform a behavior or if they could perform that behavior made more of a difference than the persuasive technology. Ultimately, adversary nations and criminal groups might attempt to victimize deployed military personnel who may be more motivated than their peers to use some persuasive technology for some purpose, such as communicating back home to family or finding love on a dating platform. But beyond those kinds of psychological underpinnings, Fogg often emphasized that health-related persuasive technologies should not being used in a coercive or deceptive manner. Some researchers have tried to demonstrate how effective persuasive technologies can be in driving positive health choices (Parmar, Keyson, and de Bont, 2009).

However, the development and application of persuasive technologies for deception and influence have largely muted Fogg's caution. Oinas-Kukkonen and Harjumaa (2008) developed a systematic framework for designing persuasive systems, finding that whether a persuasive technology is characterized as coercive or deceptive makes little difference, as peoples' attitudes and behaviors can be shaped or changed or reinforced with persuasive technology, whether deception or coercion is designed into that technology or not. Machine reasoning researcher Timotheus Kampik (2018) argued that most researchers continue to find also that the concept of deception is the root of much persuasive technology and artificial intelligence research and application, so that deception in persuasive technology not only occurs on purpose, but also "occasionally and unintentionally". Deception is generally defined as creating erroneous sensemaking in someone so that they make a decision that is advantageous to the deceiver. Kampik expanded the range of persuasive technology applications, exploring knowledge sharing websites like Stack Exchange, video sharing platforms like YouTube, and enterprise instant messaging applications like Slack and source code sharing site GitHub. In another example, Kampik analyzed subscription news media sites, such as the Washington Post online. When potential users see headlines written to attract readers (i.e. clickbait), sometimes the written content that must be subscribed to for users to read does not reflect the headline entirely. Kampik suggested this is a form of deception in the persuasive technology of that online news content site and subscription service, even if this characterization is less recognized. These applications all reflect persuasive technology design, where the design of those applications is meant to drive engagement by anticipated users based on various motivations (Kaptein et al., 2015).

In a study referred to above, Hamari, Koivisto, and Pakkanen (2014) examined a large sample of nearly 100 empirical studies on persuasive technologies and gamification. Hamari, Koivisto, and Pakkanen characterized persuasive technologies as different than gamification because there is more focus on social and communicative persuasion and attitude change, while gamification seems to concentrate more on "invoking users' (intrinsic) motivations (through gameful experiences and affordances)". Their review of these studies found that while persuasive technologies are more concentrated on changing or maintaining attitudes, only a few of these papers included general attitude as an explicit variable. Their conclusion suggested that increasingly the goal of persuasive technology design was to influence a behavior that is "valuable only for the designer of the system". This may be another indication of a shift toward more deceptive uses and motivations of persuasive technologies. The author argues that this shift will likely continue, as criminal and nation state enterprises continue to discover the design and application of persuasive technologies on targeted human behaviors.

Every person appears to have a different susceptibility as a user of persuasive technologies to the influence of whatever that technology may be (Kaptein, M., Duplinsky, S. and Markopoulos, 2011). Their research argued that the "application of personalization" in persuasive technologies could "substantially boost their impact" on someone's behaviors, whether those behaviors are related to their health or continuing to play an online game. Kaptein, Duplinksy, and Markopoulos were exploring more manual, incremental changes in user experience, calling this modification adaptive persuasive systems. Any change or modification to a technology based on a user is going to enhance the effectiveness or influence of that technology. Chen (2021) found in a study of Chinese university smartphone users that persuasive technologies increase users' screen time and contribute to the "addictive behaviors" of younger smartphone users. There is a social and individual context to the infrastructure behind these persuasive technologies. The Chinese university students needed to use their phones for financial transactions, but also for chatting with friends and family, for example. In the example of a futurist Moon Base, the author suggests that dependence on persuasive technologies will also involve similar needs. Those needs and increased use or dependence on persuasive technologies can become vulnerabilities (Wang, 2017). Khaled et al. (2006) highlighted this social and individual context as well, finding that "social software" increasingly has made group dynamics "a powerful factor" in design as a persuasive technology platform. Social software is software that enables communication and connection online. Khaled et al. noted that people generally join online platforms voluntarily, but those platforms and the communities or groups of people in relationships of some sort in those communities have established some social norms related to how people communicate or behave with each other. The use of persuasive technologies extends into these interactions among people offline, making that influence or persuasion perhaps more effective or more intense (Waardenburg, Winkel, and Lamers, 2012; Midden et al., 2011).

Spitaletta, Matthaeus, and Guadian (2018) wrote that they imagined adaptation to a user input or considerable user data to machine learning or large language models could become engines of developing persuasive technologies, dramatically driving adaptation to the user to another level or scale. The development of persuasive technologies has accelerated over the past decade due to advances in software engineering and data technologies, "from bespoke solutions in healthcare and assistive technologies to malign influence ecosystems". Technology philosopher Floridi Luciano (2024) noted that artificial intelligence likely represents another "significant leap" in the field of persuasive technologies. Luciano suggested the online persuasion of persuasive technologies identifying and shaping a user's demand and supplying the content in response to that demand or desire could be characterized as *hyper-persuasion* or *hypersuasion*. Luciano referred to Burtell and Woodside (2023) who wrote:

The relentless nature of AI's *hypersuasion*, the magnitude of its scope, its availability, affordability, and degree of efficiency based on machine-generated content accurately tailored to individual users or consumers who spend increasing amounts of their lives *onlife* (both online and offline) in the infosphere overshadow its precursors, not only in terms of the depth of personalized influence but also for the potential scale of distribution and impact.

Luciano added that artificial intelligence can and will be used more commonly and effectively to manipulate someone's views, preferences, choices, likes and displaces, hopes and fears. The author agrees that there have been growing demonstrations of persuasive technologies personalizing a user experience and that adaptation to a user experience can be intensified when backed by LLMs and other artificial intelligence generation.

3. Visualizing Scenarios of Operationalizing Persuasive Technology Design on Futurist Moon Bases

The following two theoretical scenarios visualize possible applications of persuasive technology design targeting people living and working in an outer space environment on a futurist Moon Base. The author chose a futurist Moon Base because the conditions and isolation appear to resemble futurist infrastructure and personnel access to technologies currently available in locations such as the Arctic and Antarctica. There is currently limited case studies or examination of criminal behaviors in these kinds of isolated but geopolitically and economically important environments. The visualizations were developed based on the prior cited Arctic study and more recent persuasive technology research on gamification and persuasive technology application use.

Scenario 1: "*Women's Soccer Manager Soccer Fan*"

This iPhone or Android game was developed by a Danish gaming company in approximately 2018. This game can be logistically intensive, which might be ideal if living and working in an isolated region where there is more time than normal outside of daily mission tasks or functions. Users can create rosters for their team from real

players and real soccer leagues, negotiate contracts for their players, and manage their customized team through training camp and an entire season. This season is fictional, but considerable attention would be required for maintaining the team. This game can also be downloaded onto Steam, which is a gaming platform for regular gamers where users can chat as well. There could be several people that use this gaming application who know each other socially on this futurist Moon Base, which could influence continued use of this application and maintaining a team. There could be users who have family members back on Earth who use this application as well, such as a mother of a young girl who plays soccer and follows women's soccer.

There are plausible contexts where criminals or criminal enterprises could learn through human or technical collection that someone of interest for recruitment or exploitation because of their access or role on this futurist Moon Base is spending much of their time online playing this game. That information could be sufficient for a criminal to create a controlled account or sock puppet that also plays this game and shares common interests, demonstrated in their own fictional team. Because the user playing this game may be more likely to be a woman, the controlled account would also likely appear to be a woman who may be able to connect more easily with the targeted Moon Base worker, because there are limited numbers of women in most government and military communities and even fewer deployed to isolated environments like the Arctic or this futurist Moon Base. These isolated deployment locations are likely staffed mostly with men, which could present several social challenges to the targeted female user that may make them more vulnerable to communication with someone who appears to be like them online (Cialdini and Goldstein, 2004).

This relationship development based on persuasive technology design is comprehensive, facilitating several plausible approaches for future behavioral and technical exploitation. Communicating with someone on Steam about this game is less scrutinized because the platform is public and has no association with the military or government. There also may not be much association with a criminal, because it is a platform designed for gamers or users who enjoy games.

Scenario 2: *"Duolingo Japanese Language Peer"*

This iPhone or Android free or paid language learning account is widely accessible and has become increasingly popular with users around the world. Smartphone platforms and applications that are more familiar to users are generally scrutinized less, including the users on that platform or application that a user might interact with (Cheshire et al., 2010). Civilian or military personnel from member nations who are deployed to this futurist Moon Base or even commercial representatives may be interested in learning Japanese because some of their coworkers or other people of interest to them are from Japan.

While there could be several motivations for someone to want to learn Japanese in this scenario, the persuasive technology of this language learning program, which is gamified and encourages continued interaction, maintains that behavior or encourages that behavior. There are limited opportunities to message another user on this platform, such as when a user achieves a learning badge, but that limited opportunity for contact may be sufficient to make initial contact without the user scrutinizing that contact as much (Walther et al., 2015). Praising someone who also appears to be learning the same language as you is a form of Liking and could influence continued communication with that person (Cialdini, 2001).

The outcome the targeted user may be interested in accomplishing is to be able to communicate in Japanese with his or her peers and to develop relationships with those peers. This kind of persuasive technology can facilitate that outcome and maintain or encourage behavior to achieve it. But this is also a vulnerability, because a criminal can craft a behaviorally based approach or interaction based on this persuasive technology, to build a trusting relationship or exploit that targeted user of interest in some behavioral or technical manner.

In both theoretical scenarios, the initial contact and communication occur online, from people that appear to be like the user of interest and on platforms that rely on commercial infrastructure and commercial telecommunication networks instead of government platforms. This infrastructure is generally highly vulnerable to compromise and collection, because there may be less restriction or protection compared to sensitive government or military platforms. People familiar with these applications and social platforms based on personal use generally scrutinize those platforms and the people on them less.

4. Discussion

This practitioner's paper introduced some of the foundations of persuasive technology as a conceptual framework. The paper discussed how this method of behavioral design and application for influencing and changing or maintaining attitudes and behaviors has appeared to shift toward more deceptive design and

operational uses. This shift reflects some of the author's background in law enforcement as a profiler with a federal law enforcement agency, where he behaviorally analyzed cybercriminals of interest based on their online behaviors and on what platforms or applications, they may be most behaviorally responsive to or vulnerable to an approach by someone who they may be more comfortable with or may trust more. Criminals use this same approach to socially engineer victims, but the purposeful design and application of persuasive technology to influence a victim reflects a more sophisticated approach to behaviorally exploiting someone of interest. Generative artificial intelligence run by refined LLMs and machine learning will only accelerate this approach.

This paper only referred to criminals using persuasive technology to target victims on a futurist Moon Base, but arguably this same approach could be used by intelligence and military enterprises for the same operational needs. Intelligence and military enterprises are likely already intensifying their use of this behaviorally based operational design and application method not only against people and organizations of interest online but against other nations. The author suggested that a unique environment like the Arctic which is isolated but of great geopolitical importance and interest resembles the environment on a futurist Moon Base. There are similar characterizations of those geopolitical efforts in outer space among competing nations, including which nation will establish their own infrastructure and interests on the Moon first and how they might use that foothold as leverage in resource competition and in military and political advantage. This paper suggests the people living and working on a futurist Moon base will be like the growing numbers of people living and working in the Arctic today, isolated but relying heavily on increasingly available commercial infrastructure and perhaps working in increasingly stressful conditions. Those are conditions for increased use of social platforms and communication applications.

The author visualized two scenarios where criminals might target and victimize someone of interest on a futurist Moon Base, because of their access to people or information or other material. This approach does not require the criminal to be physically located on the Moon Base as well. These futurist theoretical scenarios provided some models of what this kind of behaviorally based criminal approach using persuasive technologies might look like, highlighting this unique victimology of users in isolated environments of interest to criminals and nations.

Ethics Declaration: Ethics clearance was not required for this paper.

AI Declaration: There was no use of AI in the development or at any other stage of this paper.

References

Bashkireva, T., Bashkireva, A., Severin, A. and Morozov, A., 2019, July. Problems of psychophysiological status of human capital in the cyclic conditions of Arctic. In *IOP Conference Series: Earth and Environmental Science* (Vol. 302, No. 1, p. 012151). IOP Publishing.

Burns, R. and Lampkin, J., 2024. On the world-picture: a cross-disciplinary analysis of astronoetics and astro-green criminology. *REC. Revista Electrónica de Criminología*, *9*, pp.1-10.

Chen, X., 2021. Does Persuasive Technology Make Smartphones More Addictive?: An Empirical Study of Chinese University Students.

Cheshire, C., Antin, J., Cook, K.S. and Churchill, E., 2010. General and familiar trust in websites. *Knowledge, Technology & Policy*, *23*(3), pp.311-331.

Cialdini, R.B. and Goldstein, N.J., 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, *55*(1), pp.591-621.

Cialdini, R.B., 2001. The science of persuasion. *Scientific American*, *284*(2), pp.76-81.

Fogg, B.J., 1998, January. Persuasive computers: perspectives and research directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 225-232).

Fogg, B.J., 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity*, *2002*(December), p.2.

Fogg, B.J., 2009, April. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology* (pp. 1-7).

Hamari, J., Koivisto, J. and Pakkanen, T., 2014, May. Do persuasive technologies persuade?-a review of empirical studies. In *International conference on persuasive technology* (pp. 118-136). Cham: Springer International Publishing.

Hermida, J., 2006. Crimes in space: A legal and criminological approach to criminal acts in outer space. In *Proceedings of the Forty-Eighth Colloquium on the Law of Outer Space* (pp. 156-177).

Idrees, A.R., Kraft, R., Mutter, A., Baumeister, H., Reichert, M. and Pryss, R., 2024. Persuasive technologies design for mental and behavioral health platforms: A scoping literature review. *PLOS Digital Health*, *3*(5), p.e0000498.

Kampik, T., Nieves, J.C. and Lindgren, H., 2018. Coercion and deception in persuasive technologies. In *20th International Trust Workshop (co-located with AAMAS/IJCAI/ECAI/ICML 2018), Stockholm, Sweden, 14 July, 2018* (pp. 38-49). CEUR-WS.

Kaptein, M., Duplinsky, S. and Markopoulos, P., 2011, May. Means based adaptive persuasive systems. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 335-344).

Kaptein, M., Markopoulos, P., De Ruyter, B. and Aarts, E., 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, *77*, pp.38-51.

Khaled, R., Barr, P., Noble, J. and Biddle, R., 2006, May. Investigating social software as persuasive technology. In *International Conference on Persuasive Technology* (pp. 104-107). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lampkin, J.A. and Wyatt, T., 2025. An astro-green criminological examination of orbital space debris. *Criminology & Criminal Justice*, *25*(4), pp.1280-1297.

Lampkin, J.A. and McClanahan, B.W., 2024. Astronomical withdrawals: a green criminological examination of extreme energy mining on extraterrestrial objects. *Crime, Law and Social Change*, *81*(4), pp.365-384.

Luciano, F., 2024. Hypersuasion–On AI's persuasive power and how to deal with it. *Philosophy & Technology*, *37*(2), p.64.

Midden, C., Kimura, H., Ham, J., Nakajima, T. and Kleppe, M., 2011, June. Persuasive power in groups: the influence of group feedback and individual comparison feedback on energy consumption behavior. In *Proceedings of the 6th International Conference on Persuasive Technology: Persuasive Technology and Design: Enhancing Sustainability and Health* (pp. 1-8).

Oinas-Kukkonen, H. and Harjumaa, M., 2008, June. A systematic framework for designing and evaluating persuasive systems. In *International conference on persuasive technology* (pp. 164-176). Berlin, Heidelberg: Springer Berlin Heidelberg.

Oyibo, K., 2024. ComTech: Towards a unified taxonomy of persuasive techniques for persuasive technology design. *Computers in Human Behavior Reports*, *14*, p.100372.

Pagnini, F., Manzey, D., Rosnet, E., Ferravante, D., White, O. and Smith, N., 2023. Human behavior and performance in deep space exploration: next challenges and research gaps. *npj Microgravity*, *9*(1), p.27.

Parmar, V., Keyson, D. and De Bont, C., 2009. Persuasive technology to shape social beliefs: A case of persuasive health information systems for rural women in India. *Communications of the Association for Information Systems*, *24*(1), p.25.

Pintar, A. and Erjavec, J., 2021. A framework for designing behavioural change with the use of persuasive technology. *International Journal of Management, Knowledge and Learning*, *10*, pp.75-84.

Ramachandran, D. and Canny, J., 2008, April. Applying persuasive technologies in developing regions. In *Proceedings from ACM Conference on Human Factors in Computing Systems: Workshop on Human-Computer Interaction for Community and International Development. Florence, Italy: Association for Computing Machinery Special Interest Group on Computer-Human Interaction*.

Savishinsky, J.S., 1971. Mobility as an aspect of stress in an Arctic community 1. *American Anthropologist*, *73*(3), pp.604-618.

Sharok, V., 2020, July. Integrated assessment of the psychological characteristics of the individual who can adapt successfully to the Arctic environment. In *IOP Conference Series: Earth and Environmental Science* (Vol. 539, No. 1, p. 012110). IOP Publishing.

Spitaletta, J.A., Matthaeus, M.M.B. and Guadian, M., 2024. Cyber Considerations of a Resistance Operating Concept. *MILITARY REVIEW*, *1*.

Tariq, M.U., 2024. Harnessing persuasive technologies for enhanced learner engagement and motivation. In *Power of Persuasive Educational Technologies in Enhancing Learning* (pp. 30-62). IGI Global.

Waardenburg, T., Winkel, R. and Lamers, M.H., 2012, June. Normative social influence in persuasive technology: intensity versus effectiveness. In *International Conference on Persuasive Technology* (pp. 145-156). Berlin, Heidelberg: Springer Berlin Heidelberg.

Walther, J.B., Van Der Heide, B., Ramirez Jr, A., Burgoon, J.K. and Peña, J., 2015. Interpersonal and hyperpersonal dimensions of computer-mediated communication. *The handbook of the psychology of communication technology*, pp.1-22.

Wang, W., 2017. Smartphones as social actors? Social dispositional factors in assessing anthropomorphism. *Computers in Human Behavior*, *68*, pp.334-344.

How Ends, Ways, and Means are Manifested in the Cyber Defense Strategies of Superpowers

Piia Perälä, Martti Lehto and Pekka Pirinen
Faculty of Information Technology, University of Jyväskylä, Finland

piia.m.h.perala@jyu.fi
martti.j.lehto@jyu.fi
pekka.j.pirinen@jyu.fi

Abstract: Cyber Power refers to the capacity to project and promote national interests in and through cyberspace. The military contributes to multiple aspects of cyber power, with cyber warfare capabilities being a central component. Numerous state and non-state actors have come to regard cyber ends, ways and means, as a powerful force multiplier, essential to achieving their objectives. Military cyber actors employ malicious cyber operations to gain asymmetric advantages, targeting critical infrastructure and undermining the opponent's military superiority in cyberspace. The development of cyber warfare has seen a convergence with other non-kinetic warfare operations. Superpowers adopt distinct approaches to cyber warfare, with varying definitions, strategic objectives, and methods for achieving them. They have also established their own command structures and cyber forces. This paper aims to provide an understanding of how superpowers use cyber capabilities and leverage asymmetric advantages to achieve their strategic objectives. The research material consisted of publicly available documents on national strategic cyber defense plans. Lykke's framework was applied as a structured approach to review superpowers' strategic plans by identifying the desired outcomes (Ends), the methods used to achieve them (Ways), and the resources required (Means).

Keywords: Superpowers, Cyber warfare, Ends, Ways, Means, Convergence, Cyber defense strategy, Cyber power

1. Introduction

Cyber power is the ability of a state to advance national interests through cyberspace. Military cyber capabilities form a core component of this power, as cyber warfare has become a critical tool for achieving strategic objectives. (Voo *et al.*, 2020) States and non-state actors increasingly view cyber operations as force multipliers, enabling asymmetric advantages. Cyber warfare has converged with other non-kinetic domains, challenging traditional concepts of conflict (Lehto and Limnéll, 2017). Nations have developed distinct approaches to cyber warfare, shaped by their strategic cultures, political systems, and technological capabilities. Definitions of cyber warfare and its objectives vary across nations, as do the ways to achieve goals (Giles, 2021). While some nations prioritize technical dominance and infrastructure resilience, others emphasize cognitive influence and information control.

This paper aims to provide an understanding of how the cyber superpowers, namely the United States (US), Russia, and China, employ cyber capabilities to achieve strategic objectives. Applying Lykke's framework, we analyzed publicly available documents outlining these nations' strategic cyber defense plans, aiming to identify their cyber defense ends, ways, and means. The paper provides insight into how these superpowers leverage cyber capabilities and asymmetric advantages to accomplish their strategic objectives.

2. Cyber Warfare, Power, and Deterrence

2.1 Cyber Warfare

Since the 1990s, cyber warfare has emerged as a parallel domain to traditional kinetic warfare, often complementing conventional military operations (Haizler, 2017). It is a key element of hybrid warfare, which combines military force, irregular tactics, and criminal activities by state or non-state actors (Lalu and Puistola, 2015). Cyber warfare has challenged the traditional paradigm of warfare. Operating in the cyber environment has reduced the relevance of formal declarations of war or peace. In cyberspace, conflicts lack clear beginnings or endings, resulting in prolonged periods of instability for targeted entities. (Lehto and Limnéll, 2017)

Defining cyber warfare is challenging, as no universally accepted definition exists (Haizler, 2017). The term broadly covers diverse activities in the cyber domain, yet certain common characteristics can still be identified. Cyber warfare operates in a borderless digital battlespace with no front lines, exposing both military and civilian systems. Its rapid tempo shifts conflicts to seconds, and its impact reaches critical infrastructure and public services. Cyber-attacks target both virtual systems and physical hardware, aiming for systemic disruption rather than isolated damage. Often integrated with kinetic and hybrid warfare, these operations combine cyber, information, and electronic tactics. Cyber operations can occur without formal declarations of war, typically by

state or state-sponsored actors, making cyber threats persistent and low-threshold in modern conflict. (Lehto and Limnéll, 2017)

2.2 Cyber Power

In the modern world, cyber capabilities have become a key element of national power (IISS, 2021). Cyber power enables states to achieve strategic objectives through cyberspace, with military applications playing a central role (Voo *et al.*, 2020). Definitions converge on the idea that cyber power is a state's ability to leverage cyberspace to influence outcomes and pursue goals (Kuehl, 2009; Nye, 2010; Voo *et al.*, 2020). Cyber power is shaped by technology, organizational structures, and user capabilities. Its development and application are influenced by national missions, such as military, economic, or political (Kuehl, 2009). It can affect both digital and physical domains and serve strategic or political aims (Lehto and Limnéll, 2017). Unlike traditional power, cyber power is not dependent on a state's size or proximity. Instead, it hinges on clear objectives, the means to achieve them, and leveraging cyber means to accomplish multiple objectives. (Lehto and Limnéll, 2017; Voo *et al.*, 2020).

2.3 Cyber Deterrence

While cyber deterrence typically operates within the cyber domain, its effects can extend beyond it (Lonergan and Montgomery, 2021). Effective deterrence depends on a credible threat that convinces potential attackers the costs outweigh the benefits. States adopting this strategy may not seek conflict but remain prepared to defend vital interests if necessary (Pijpers and Arnold, 2025). According to Lehto and Limnéll (2017), building cyber deterrence requires the ability to defend society broadly in the cyber environment, tolerate attacks effectively, and attribute the source of those attacks. States must also be capable of launching counterattacks and responding through political means when needed. A sufficient level of cyber self-sufficiency and competence is essential. Public communication strengthens cyber deterrence. Although cyber capabilities are often classified, states may disclose or demonstrate certain capabilities to enhance credibility (Limnéll, 2013).

3. The US, Chinese, and Russian Approaches to Warfare in Cyberspace

The US military defines cyberspace as “a global domain within the information environment consisting of the interdependent networks of information technology infrastructures and resident data, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers” (US DoD, 2021, p. 55). Cyber operations involve using these capabilities to achieve objectives within or through cyberspace, while cyber warfare refers to armed conflict conducted partly or wholly via cyber means. Its goal is to deny adversaries the effective use of cyberspace systems and includes attacks, defense, and enabling actions (US DoD, 2010). Unlike the US's technical and infrastructure-focused view, China and Russia adopt broader approaches. China views the information space as a global communication domain, aligning with Russia’s concept of a cognitive space involving human information processing (Giles and Ii, 2013).

China treats information as a distinct domain of warfare, equal to land, sea, air, space, and cyber, with a strong focus on ideological control and perception management (Giles and Ii, 2013). Its concept of information spans cyberspace, the electromagnetic spectrum, physical infrastructure, and the cognitive domain (Mulvaney, 2025). China emphasizes the close link between space and cyber, viewing both as extensions of the electromagnetic spectrum (Costello and McReynolds, 2018). China perceives peace and war as a continuum, not separate states, and integrates offensive cyber operations into a broader strategy of information warfare that spans military, diplomatic, and economic spheres. Like Russia, China combines cyber capabilities and information warfare, emphasizing the fusion of multiple warfare types utilizing information technologies. Its cyber strategy is tied to social stability and military modernization. (Pusztaszeri, Harding and Dickson, 2025) The development of an ‘informatized’ and ‘intelligentized’ military through ‘digitization’, ‘networkization’, and AI-driven smartification is central to the evolution of the People's Liberation Army (PLA) doctrine and future conflict readiness (Mulvaney, 2025).

Russia considers cyber and information operations inseparable, framing its activities under the concept of ‘information confrontation’. This approach encompasses a broader and continuous range of actions than the US notion of cyberwar, extending beyond wartime operations (Dickson and Harding, 2025). Information confrontation includes both technical and psychological dimensions, with cognitive operations playing a central role in shaping public perception and influencing decision-making (Hakala and Melnychuk, 2021). Russia avoids the term ‘cyber warfare’ and, together with China, shares the understanding that it describes foreign concepts, mainly Western threats and activities (Giles and Ii, 2013).

4. Lykkes' Framework

In 1989, Colonel Arthur F. Lykke Jr. presented a model concerning military strategy. Lykke's (Lykke, 1989) model suggests that military strategy, or any other strategy (e.g., political, economic etc.), can be viewed as an equation in which ends, ways, and means form the strategy. Ends refer to the objectives to be achieved. Ways describe the actions that need to be taken to reach the goals, and means are the resources required to achieve the ends. These strategic concepts support approaches at all levels: strategic, operational and tactical. Applying Lykke's model to cyber defense strategies, 'ends' refer to strategic objectives such as defending against cyber threats. 'Ways' refers to the methods used (e.g., disrupting malicious actors), and 'Means' refers to the required resources (e.g., skilled personnel, funding, and advanced cyber tools).

5. Method

The research material consisted of publicly available documents outlining strategic cyber defense plans of the US, Russia, and China. The appropriateness of the documents was evaluated against four criteria (Scott, 2014): authenticity, credibility, representativeness, and meaning. For example, we reviewed document details (e.g., authors, publishers, affiliated organizations, and publication year) to ensure authenticity and confirm that the documents were from credible sources. We also ensured that the content was relevant to our study aims. Only documents written in languages understood by the authors were included.

For the analysis, we used thematic analysis (Braun and Clarke, 2012) and applied Lykke's (1989) framework to identify key concepts: ends, ways, and means. We began the analysis by familiarizing ourselves with the research materials, after which the data was coded and features corresponding to Lykke's concepts were identified. We utilized Atlas.ti for coding. Codes that shared common characteristics were clustered to generate themes and sub-themes.

6. Ends, Ways, and Means of Superpowers in Cyber Defense

6.1 United States

6.1.1 Strategic ends and corresponding ways of the US

End: Dominance in cyberspace. The US aims to achieve and maintain cyberspace superiority in both peace and conflict, using cyber operations to secure lasting advantages and enhance its kinetic and informational strengths (USCYBERCOM, 2018; IISS, 2021).

Ways: The US is building capabilities that support and enable the full spectrum of activities in cyberspace. The US conducts diverse operations that enhance national power and support the Joint Force, including persistent campaigning, cyber defense, resilience-building, and strategic planning. (DoD, 2023). It engages in detailed planning and ensures the broad dissemination of comprehensive cyber strategies. As a global leader in space connectivity, the US has effective military cyber operations for intelligence, damage assessment, and targeting. (IISS, 2021) The US also has a leading role in global cyber alliances and communities, having effectively defended its critical infrastructure and promoted shared security principles. Successful diplomacy, leadership in organizations (e.g., IEEE and ISACA), and contributions to international technical standards strengthen its dominance in cyberspace. (IISS, 2021)

End: Cautious approach of operations. The US aims to manage the risk of unintended escalation and carefully weigh whether to exploit a vulnerability for disruption or preserve access for ongoing intelligence collection. (DoD, 2023; Harding, Pusztaszeri and Dickson, 2025).

Ways: The US conducts cyber operations below the level of armed conflicts and to achieve favorable security (DoD, 2023). The US patiently plans and constructs a concept for the operation, maps the target systems, and waits for third-party approval, but does not proceed further until an opportunity presents itself (Harding, Pusztaszeri and Dickson, 2025).

End: Strong alliances and partners. The US views its global alliances as a core strategic advantage in cyberspace (DoD, 2023).

Ways: The US strengthens ties with cyber-capable allies at the strategic, operational, and tactical levels, integrating these efforts into broader security cooperation (DoD, 2023). Collaboration with NATO, Pacific allies, and Five Eyes partners enhances US cyber operations through joint incident response, shared attribution, and coordinated actions against malicious actors (IISS, 2021; Harding, Pusztaszeri and Dickson, 2025). Through 'defend forward' operations, the US Cyber Command disrupts adversary activities in coordination with allies.

The US builds allied cyber capacity by sharing infrastructure, conducting joint training, providing specialized capabilities, and identifying partner network vulnerabilities to enhance collective cybersecurity and readiness. (DoD, 2023) It leads initiatives (e.g., Cyber Deterrence Initiative), building coalitions to deter malicious cyber activity through shared intelligence, joint attribution, public support, and collective penalties. (Nye, 2010). The US adopts a unified cybersecurity approach, integrating government, industry, academia, state authorities, defense, the National Guard, and privacy advocates (IISS, 2021). The private sector is a vital role in analyzing threats and protecting national infrastructure (Harding, Pusztaszeri and Dickson, 2025).

End: Deterrence building. Cyberspace operations are essential to US military power, supporting integrated deterrence by delivering informational and strategic advantages (DoD, 2023).

Ways: The US conducts cyberspace campaigns that integrate cyber operations with other intelligence disciplines to gather threat intelligence. It uses cyber espionage to monitor leadership decisions, tactical planning, and foreign governments, while also deploying cyber tools to disrupt adversary information operations and nuclear programs, including retaliatory strikes on foreign systems and infrastructure. The US uses deterrence through punishment by responding to hostile acts with great force, making such actions costly for adversaries. It also deters through denial by deploying defenses measures that limit attack impact and discourage repetition. (Harding, Pusztaszeri and Dickson, 2025)

End: 'Defend forward' approach. The US follows a proactive strategy known as 'defend forward', which involves disrupting malicious cyber actors before they can launch attacks. The US is committed to defending its national interests in cyberspace by maintaining readiness to protect its citizens, critical infrastructure, and territorial integrity (IISS, 2021; DoD, 2023; Harding, Pusztaszeri and Dickson, 2025). The US leverages cyberspace as a strategic domain to deter aggression, maintain readiness, and win wars (DoD, 2023).

Ways: The US cyber strategy aligns with national objectives and prioritizes resilient defenses built on strong infrastructure, strategic operations, and advanced technologies (Harding, Pusztaszeri and Dickson, 2025). The US enhances its cyber capabilities to support the Joint Force's mission assurance and collaborates with the science and technology community to develop tools that disrupt malicious cyber actors. The US trains its forces to operate in degraded cyber environments and learns from both adversary and internal operations to improve cyber warfare capabilities and risk management. Acknowledging that cyber risks span the entire defense enterprise, the US strengthens cybersecurity culture across defense by educating leaders, integrating training, and investing in a diverse workforce supported by incentives and private-sector collaboration. (DoD, 2023)

End: Integrating cyber operations with military operations. The US integrates cyber operations into military planning to gain strategic and asymmetric advantages.

Ways: The US develops cyberattack capabilities for all operational phases and across command levels, with each military service maintaining cyber units for offensive and defensive missions. These units support combatant commands and conduct offensive cyber operations when directed (Harding, Pusztaszeri and Dickson, 2025). As part of integrated deterrence, cyber operations are embedded in campaign and contingency planning. The US leverages cyberspace's unique characteristics to achieve Joint Force requirements, including cross-domain effects in large-scale combat operations. (DoD, 2023)

6.1.2 Identified means

The US aims for dominance and deterrence in cyberspace through advanced infrastructure, persistent operations, and integrated intelligence. Strategic planning, legal oversight, and decision-support systems guide cautious execution. Alliances are strengthened via joint training, shared infrastructure, and coalition coordination, while private sector collaboration boosts threat analysis and infrastructure defense. Deterrence relies on offensive tools, espionage, and strong defenses. The 'defend forward' approach uses proactive disruption, resilient systems, and a skilled, diverse workforce. Integration with military operations requires specialized cyber units, cross-domain platforms, and campaign planning to support Joint Force missions.

6.2 Russia

6.2.1 Strategic ends and corresponding ways of Russia

End: Dominance in cyberspace. Russia seeks greater cyber dominance as part of its effort to reclaim great power status (Voo and Singh, 2025).

Ways: Russia emphasizes information as a pillar of state power and uses it as a key tool of statecraft, leveraging cyber operations, propaganda, and strategic messaging to assert power, shape domestic narratives, influence

the global order, and revise the European security architecture. It emphasizes a comprehensive approach to controlling and deploying information both domestically and internationally. (Voo and Singh, 2025) Russia views offensive cyber operations as a key means of achieving information dominance across all stages of conflict (Connell and Vogler, 2017).

End: 'Information confrontation' approach. Russia views the internet as a persistent threat and a strategic weapon, treating the information space as a continuous battlefield where cyber and psychological operations advance long-term objectives (Connell and Vogler, 2017; Sherman, 2025). The 'information confrontation' strategy merges cyber and information operations, encompassing espionage, infrastructure attacks, and psychological manipulation (Dickson and Harding, 2025).

Ways: Russia's belief that it is engaged in an information war with the West drives a proactive and defensive strategy focused on controlling domestic narratives and influencing foreign targets (Dickson and Harding, 2025). Information confrontation is regarded as a strategic tool that can substitute for conventional military force under certain conditions. (Dickson and Harding, 2025). Russia's Information Confrontation Doctrine is supported by diverse funding sources (e.g., the Ministry of Defense and other security bodies), which sustain cyber units and offensive cyber operations (Voo and Singh, 2025). Russia is enhancing its tech-savvy information warfare capabilities, leveraging cybercriminal proxies, deepfakes, and generative AI to increase campaign scale and complexity. Most operations target societal and governmental stability rather than military systems. (Voo & Singh, 2025; Dickson & Harding, 2025)

End: Cooperation. Russia cooperates within its cyber ecosystem to enable actions in cyberspace.

Ways: Russia's cyber ecosystem integrates state agencies, cybercriminals, contractors, patriotic hackers, and private military companies (Sherman, 2025). Russia sustains relationships with cybercriminal groups, which it tolerates and exploits under an informal "social contract" requiring them to target foreign entities, avoid interfering with state objectives, and remain responsive to government directives. (Dickson and Harding, 2025; Sherman, 2025). These actors provide services (e.g., DDoS attacks, malware obfuscation, etc.) while generating revenue and cultivating cyber talent for state use (Connell and Vogler, 2017). Russian intelligence agencies recruit hackers and contractors for offensive operations, including espionage, sabotage, and influence campaigns. Cybercriminal tools have also been adapted for state purposes (e.g. NotPetya attack) (Voo and Singh, 2025).

End: Strategic deterrence. Russia is increasingly using offensive cyber operations to support conventional military actions and integrate them into its strategic deterrence (Connell and Vogler, 2017).

Ways: Strategic deterrence involves coordinated forceful and non-forceful measures to prevent adversaries from taking actions that could cause harm (Dickson and Harding, 2025). Russia employs cyber operations, disinformation, and covert influence to intimidate neighboring states and discourage alignment with Western institutions (e.g., NATO and the EU), while asserting regional dominance and countering Western integration (Voo and Singh, 2025). It also uses information and psychological operations to incite fear and instability in adversary populations, staying below the war threshold to avoid retaliation (Dickson and Harding, 2025).

End: Cyber readiness. Russia's cyber operations fall into two types: high-profile actions for plausible deniability and covert campaigns for espionage and future attacks. State actors provide stealth, cybercriminals offer deniability, and IT firms supply expertise and R&D. (Voo and Singh, 2025)

Ways: Russia's strategy combines espionage, disruption, and influence, targeting infrastructure and shaping public perception, while psychological tactics aim to manipulate opinion and distract governments (Lilly and Cheravitch, 2020; Dickson and Harding, 2025). A decentralized network of proxies, hacktivists, and cybercriminals, often linked to intelligence services, enables disinformation and disruptive campaigns, while avoiding direct attribution. Ambiguity is reinforced through inconsistent control of these actors and the use of front organizations to obscure state involvement. (Connell and Vogler, 2017; Lilly and Cheravitch, 2020; Dickson and Harding, 2025) Russia's cyber strength is built on a legacy of Soviet engineering education and a strong talent pool, drawing recruits from both legal and criminal sectors. It is supported by a hybrid culture that merges digital expertise with military strategy. (Connell and Vogler, 2017; Soldatov and Borogan, 2022)

End: Integration of military, cyber, and information operations. Russia considers cyber and information operations central to its military and intelligence strategy (Dickson and Harding, 2025; Sherman, 2025).

Ways: Russia's concept of warfare has evolved to include non-military measures, such as psychological, electronic, and cyber operations, that can be deployed before or without conventional conflict. These

capabilities enable covert influence over infrastructure and populations, blurring the line between peace and war. (Lilly and Cheravitch, 2020; Voo and Singh, 2025) Although Russia stresses defense, its actions indicate an increasing focus on offensive capabilities (Connell and Vogler, 2017).

End: Information influence. Shaping adversary perceptions is central to Russia's strategy.

Ways: Russia ensures regime stability by controlling its domestic information space through censorship, surveillance, infrastructure manipulation, and suppression of dissent, reinforcing narratives of external threats. Technical and psychological tools (e.g., cyber capabilities, legislation, media control, and strategic messaging) are used to suppress opposition, unify public opinion, and counter Western influence. (Voo and Singh, 2025) Russia combines infrastructure control and cognitive manipulation to influence adversary perceptions and decisions (Lilly and Cheravitch, 2020). It conducts global information warfare, especially in Africa, Latin America, and the Middle East, using cyber operations, disinformation, military presence, and covert funding to pursue geopolitical goals. Domestically and abroad, Russia relies on state media, education, historical revisionism, and influence campaigns to dominate the information space. It employs sophisticated narrative manipulation, using contradictory messaging and reflexive control to confuse, destabilize, and steer adversaries toward favorable outcomes. (Voo and Singh, 2025)

6.2.2 Identified means

Russia's pursuit of cyber and information dominance relies on extensive resources, including state institutions and intelligence agencies, a robust cyber talent pool drawn from both legal and criminal sectors, and a legacy of strong technical education. It leverages cybercriminal groups, patriotic hackers, private military companies, and IT firms to provide technical expertise, operational capabilities, and plausible deniability. Advanced tools enhance the sophistication of operations, while state media, educational systems, and legal mechanisms enable control over domestic narratives and global influence campaigns.

6.3 China

6.3.1 Strategic ends and corresponding ways of China

End: Information dominance. China views information dominance as central to modern warfare, enabling effective operations across all domains (air, land, sea, space, cyberspace, and the electromagnetic spectrum).

Ways: Controlling information is considered critical for managing escalation and gaining strategic initiative (Cheng, 2019). China emphasizes denying adversaries access to information domain to achieve operational superiority (Costello and McReynolds, 2018). Its military strategy focuses on targeting enemy networks, command systems, and combat units through information warfare, including influence operations and precision strikes on IT infrastructure (Cheng, 2019). China also seeks to degrade adversary decision-making by disrupting information processing, transmission, and collection through both kinetic and non-kinetic strikes. A key concept is 'information isolation', which involves disabling C4ISR systems to reduce situational awareness and delay responses. (Burke *et al.*, 2020)

End: Paralysis and cognitive dominance. China prioritizes disabling an adversary's systems through targeted paralysis and decapitation rather than outright destruction.

Ways: China aims to "win without fighting" by shaping adversary decisions and achieving objectives below open warfare. This includes, for example, rapidly dominating and resolving conflicts, providing strategic information support, and emphasizing the importance of preemption in information warfare. (Costello and McReynolds, 2018; IISS, 2021) China's 'Three Warfares' strategy encompasses psychological, public opinion, and legal warfare. It regards the cognitive space as a critical battlefield. The strategy seeks to shape narratives, weaken adversaries' resolve, and secure diplomatic and informational advantages through psychological impact, particularly in the early stages of a conflict. (Burke *et al.*, 2020)

End: Cooperation. China's cyber strategy centers on cooperation with military, civilian, and private actors to build a state-supported cyber ecosystem and gain asymmetric advantages.

Ways: Through military-civil fusion, China partners with tech firms to enhance capabilities and global digital influence. It employs contract hackers, state-sponsored groups (e.g., Volt Typhoon), and hacktivists for espionage, infrastructure infiltration, and intellectual property theft, often blurring the lines between state and non-state actors (IISS, 2021; Pusztaszeri, Harding and Dickson, 2025). China's cybersecurity infrastructure includes national research centers and commissions that integrate private sector expertise into military cyber operations (Pusztaszeri, Harding and Dickson, 2025). Internationally, China promotes cyber norms emphasizing

state control, working closely with Russia and SCO members, and supports broader state participation through the OEWG to counter Western influence (IISS, 2021).

End: Active defense. China follows an 'Active Defense' strategy, asserting that it will not initiate cyberattacks but will retaliate if attacked (Cheng, 2019; Pusztaszeri, Harding and Dickson, 2025).

Ways: China focuses on strengthening cyber defenses to ensure resilience and effective response. To protect cyberspace sovereignty, it employs economic, legal, diplomatic, and military tools, including pre-positioned assets for precision "hard kill" attacks such as disabling power grids. (Pusztaszeri, Harding and Dickson, 2025) China's cyber operations prioritize espionage and information exploitation to advance political and economic objectives, including intelligence gathering for strategic advantage, AI development, and cyber readiness. (IISS, 2021; Pusztaszeri, Harding and Dickson, 2025) China integrates espionage with offensive capabilities to enable flexible responses below escalation thresholds, targeting intellectual property, personal data, and critical infrastructure (Costello and McReynolds, 2018; IISS, 2021).

China's cyber strategy aligns with national goals. It combines targeted and opportunistic attacks and includes malware and disinformation cooperation with Russia (Pusztaszeri, Harding and Dickson, 2025). China emphasizes speed and precision, accelerating the reconnaissance-control-attack-evaluation cycle, with PLA commanders acting on timely intelligence (Burke *et al.*, 2020). Advanced C4ISR technologies support information superiority in joint operations. (Pusztaszeri, Harding and Dickson, 2025). China aims to reduce dependence on foreign technology and lead in core internet technologies through initiatives (e.g., Made in China 2025 and the Digital Silk Road), promoting Chinese infrastructure and standards globally. It shapes global standards in emerging technologies (e.g., IoT, IPv6), pursues AI leadership, and develops space and cyber forces to strengthen cyber power. (Cheng, 2019; IISS, 2021; Pusztaszeri, Harding and Dickson, 2025)

End: Integrating cyber and military operations. China is modernizing the PLA to conduct informatized operations under its "systems destruction warfare" strategy (Burke *et al.*, 2020).

Ways: The "systems destruction warfare" strategy supports political objectives by gaining information dominance, expanding strategic reach, and dismantling adversary systems (Burke *et al.*, 2020). The PLA prioritizes real-time data sharing, precision strikes, and intelligent munitions to enable low-collateral, integrated warfare across physical and informational domains. Early operations focus on shaping the battlespace through intelligence and non-kinetic actions, delaying strikes until information superiority is achieved. Joint operations integrate all service branches through networked, digitized systems, breaking hierarchies and synchronizing efforts across levels, with emphasis on cyber, electromagnetic, and space domains. (Burke *et al.*, 2020; IISS, 2021)

End: Control and censorship system. China regards cyberspace as a domain with weak international norms, asserting sovereignty over its own internet while exploiting access to others. This asymmetry legitimizes censorship, media manipulation, and cyberattacks (including on civilian infrastructure), as strategic tools. (Pusztaszeri, Harding and Dickson, 2025)

Ways: Domestically, China prioritizes thought control and suppression of Western ideologies to maintain stability, which it considers essential for national development and public well-being. Foreign cyber interference (e.g., political disruption or data theft) is considered a major threat to internal security and personal data protection. (Pusztaszeri, Harding and Dickson, 2025) China operates the world's most advanced cyber surveillance and censorship system. The Great Firewall restricts foreign content and limits external access to Chinese platforms, reinforcing information dominance and shaping public perception. Information warfare combines content restrictions with economic and cyber pressure to deter foreign influence. The Ministry of Public Security manages domestic surveillance and cybersecurity. It conducts inspections, accesses user data, and evaluates system vulnerabilities, impacting nearly all foreign firms and strengthening state control over corporate and personal information. (Pusztaszeri, Harding and Dickson, 2025) China's laws require foreign firms to store data locally, increasing state control. China's cyber governance model influences other authoritarian states (e.g., Vietnam and Russia), which are adopting similar internet regulations. Additionally, China exports surveillance technologies (e.g., facial and voice recognition and smart city systems) that enable mass data collection and AI-driven monitoring. (IISS, 2021)

6.3.2 Identified means

China invests in advanced C4ISR systems, AI tools, and space-based ISR platforms to achieve information dominance. For paralysis and cognitive dominance, it needs psychological warfare teams and information-

influence capabilities. China builds cooperation through partnerships with tech firms, national research centers, state-sponsored hackers, and international alliances. Its active defense strategy involves robust cyber defenses, pre-positioned strike assets, and intelligence systems. Integration of cyber and military operations requires modernized PLA units, networked command systems, intelligent munitions, and space-cyber forces. The control and censorship system relies on the Great Firewall, AI-driven surveillance technologies, strict data localization laws, Ministry of Public Security infrastructure, and exportable monitoring tools to reinforce domestic control and influence abroad.

7. Discussion and Conclusion

Cyber superpowers share similar high-level strategic goals related to dominance, approach, cooperation, and integration. They also pursue goals concerning defense, deterrence, influence, readiness, and control, which vary by nation. Although nations may share similar strategic goals, the measures taken to achieve them differ, as the US, Russia, and China pursue cyber power through distinct strategic approaches.

The US emphasizes defense, deterrence, and cooperation, integrating cyber capabilities into conventional military planning and global alliances. It maintains a cautious, rules-based approach. Russia adopts a confrontational and disruptive model, viewing cyberspace as a battlefield for psychological and informational warfare. It leverages a decentralized ecosystem of proxies and cybercriminals to achieve strategic objectives, often targeting societal stability and public perception. China aims for comprehensive information dominance, integrating cyber operations deeply into military modernization and global influence campaigns. Its strategy combines state control, technological development, and cognitive warfare, with the goal of shaping both domestic and international narratives.

This paper has several limitations. The material used in the analysis consists of Western literature, which may introduce a cultural or geopolitical bias in the analysis and interpretation. Additionally, due to language constraints, the authors relied on literature available in languages they could understand, potentially excluding relevant works published in other languages. Future research could address a broader range of original documents and adopt a more systematic approach to selecting source material. Expanding the review to include source material from countries known for their advanced cyber capabilities would provide a more diverse perspective on the topic.

AI declaration: AI tools were not used in the creation of this paper.

Ethics declaration: An ethics declaration was not required for the research.

References

Braun, V. and Clarke, V. (2012) "Thematic analysis," in H. Cooper, P.M. Camic, D.L. Long, A.T. Panter, D. Rindskopf, and K.J. Sher (eds.) *APA handbook of research methods in psychology*. American Psychological Association (Research designs: Quantitative, qualitative, neuropsychological, and biological, 2), pp. 57–71. Available at: https://doi.org/10.1037/13620-004.

Burke, E.J., Gunness, K., Cooper, C.A.I. and Cozad, M. (2020) *People's Liberation Army Operational Concepts*. RAND Corporation. Available at: https://www.rand.org/pubs/research_reports/RRA394-1.html (Accessed: November 10, 2025).

Cheng, D. (2019) *Testimony before U.S.–China Economic and Security Review Commission*. Heritage Foundation. Available at: https://www.uscc.gov/sites/default/files/Cheng_USCC%20Testimony_FINAL.pdf (Accessed: November 10, 2025).

Connell, M. and Vogler, S. (2017) *Russia's Approach to Cyber Warfare*. Center for Naval Analyses.

Costello, J. and McReynolds, J. (2018) "China's Strategic Support Force: A Force for a New Era." National Defense University Press Washington, D.C. Available at: https://digitalcommons.ndu.edu/china-strategic-perspectives/6.

Dickson, J. and Harding, E. (2025) *A Playbook for Winning the Cyber War, Part 2: Evaluating Russia's Cyber Strategy*. Center for Strategic and International Studies (CSIS). Available at: https://csis-website-prod.s3.amazonaws.com/s3fs-public/2025-09/250903_Dickson_Playbook_Russia.pdf?VersionId=6RPe2gNRM05DWfEZqmvcyyukhE28aVbk (Accessed: November 11, 2025).

DoD (2023) "2023 DOD Cyber Strategy Summary." U.S. Department of Defense. Available at: https://media.defense.gov/2023/Sep/12/2003299076/-1/-1/1/2023_DOD_Cyber_Strategy_Summary.pdf (Accessed: November 9, 2025).

Giles, K. (2021) "Russian information warfare," in T. Clack and R. Johnson (eds.) *The World Information War*. 1st ed. Abingdon: Routledge, pp. 139–161. Available at: https://doi.org/10.4324/9781003046905-12.

Giles, K. and Ii, W.H. (2013) "Divided by a Common Language: Cyber Definitions in Chinese, Russian and English," in. *2013 5th International Conference on Cyber Conflict (CYCON 2013)*, IEEE, pp. 1–17.

Haizler, O. (2017) "The United States' cyber warfare history: Implications on modern cyber operational structures and policymaking," *Cyber, Intelligence, and Security*, 1(1), pp. 31–45.

Hakala, J. and Melnychuk, J. (2021) *Russia's Strategy in Cyberspace*. Riga: NATO Strategic Communications Centre of Excellence. Available at: https://stratcomcoe.org/pdfjs/?file=/publications/download/Nato-Cyber-Report_15-06-2021.pdf?zoom=page-fit (Accessed: October 22, 2025).

Harding, E., Pusztaszeri, A. and Dickson, J. (2025) "A Playbook for Winning the Cyber War: Part 5: Evaluating U.S. Cyber Strategy." Available at: https://www.csis.org/analysis/playbook-winning-cyber-war-part-5-evaluating-us-cyber-strategy (Accessed: November 9, 2025).

IISS (2021) *Cyber Capabilities and National Power: A Net Assessment*. International Institute for Strategic Studies. Available at: https://www.iiss.org/research-paper/2021/06/cyber-capabilities-national-power/ (Accessed: October 25, 2025).

Kuehl, D. (2009) "From Cyberspace to Cyberpower: Defining the Problem," in F.D. Kramer, S.H. Starr, and L.K. Wentz (eds.) *Cyberpower and National Security*. University of Nebraska Press, pp. 24–42. Available at: https://doi.org/10.2307/j.ctt1djmhj1.

Lalu, P. and Puistola, J. (2015) *Hybridisodankäynnin käsitteestä*. Research review 1/2015. Riihimäki: Puolustusvoimien tutkimuslaitos.

Lehto, M. and Limnéll, J. (2017) "Kybersodankaynnin kehityksesta ja tulevaisuudesta," *Tiede ja ase*.

Lilly, B. and Cheravitch, J. (2020) "The Past, Present, and Future of Russia's Cyber Strategy and Forces," in *2020 12th International Conference on Cyber Conflict (CyCon)*. *2020 12th International Conference on Cyber Conflict (CyCon)*, Estonia: IEEE, pp. 129–155.

Limnéll, J. (2013) "Offensive cyber capabilities are needed because of deterrence," in J. Rantapelkonen and M. Salminen (eds.) *The fog of cyber defence*. National Defence University/Department of Leadership and Military Pedagogy (2), pp. 200–207.

Lonergan, E. and Montgomery, M. (2021) "What is the Future of Cyber Deterrence?," *SAIS Review of International Affairs*, 41(2), pp. 61–73.

Lykke, A.F. (1989) "Defining military strategy," *Military Review*, 69(5), pp. 2–8.

Mulvaney, D.B.S. (2025) *PLA Views on the Information Domain*. China Aerospace Studies Institute.

Nye, J. (2010) *Cyber Power*. Belfer Center for Science and International Affairs, Harvard Kennedy School. Available at: https://www.belfercenter.org/sites/default/files/pantheon_files/files/publication/cyber-power.pdf (Accessed: October 25, 2025).

Pijpers, P. and Arnold, K. (2025) "Rethinking Cyber Deterrence: Adapting to the Realities of the Digital Battlefield," *Journal of Strategic Security*, 18(1), pp. 61–76.

Pusztaszeri, A., Harding, E. and Dickson, J. (2025) *A Playbook for Winning the Cyber War, Part 3: Evaluating China's Cyber Strategy*. A Report of the CSIS Intelligence, National Security, and Technology Program. Center for Strategic and International Studies (CSIS). Available at: https://www.csis.org/analysis/playbook-winning-cyber-war-part-3-evaluating-chinas-cyber-strategy (Accessed: October 11, 2025).

Scott, J. (2014) *A Matter of Record: Documentary Sources in Social Research*. John Wiley & Sons.

Sherman, J. (2025) *Confronting Russia's cyber power: reassessing assumptions, sizing up the threat, and building a proactive response*. Washington, DC: Atlantic Council.

Soldatov, A. and Borogan, I. (2022) *Unpacking the Kremlin's Capabilities*. Center for European Policy Analysis. Available at: https://cepa.org/wp-content/uploads/2022/09/Unpacking-Russian-Cyber-Operations-9.2.22-1.pdf (Accessed: November 10, 2025).

US DoD (2010) "Joint Terminology for Cyberspace Operations." US Department of Defense.

US DoD (2021) "DOD Dictionary of Military and Associated Terms." Department of Defense.

USCYBERCOM (2018) "Achieve and Maintain Cyberspace Superiority: Command Vision for US Cyber Command." The US Cyber Command. Available at: https://www.cybercom.mil/Portals/56/Documents/USCYBERCOM%20Vision%20April%202018.pdf (Accessed: November 9, 2025).

Voo, J., Hemani, I., Jones, S., DeSombre, W., Cassidy, D. and Schwarzenbach, A. (2020) *Reconceptualizing Cyber Power*. Harvard Kennedy School Belfer Center for Science and International Affairs. Available at: https://www.belfercenter.org/publication/reconceptualizing-cyber-power (Accessed: October 1, 2025).

Voo, J. and Singh, V.V. (2025) *Russia's Information Confrontation Doctrine in Practice (2014–Present): Intent, Evolution and Implications*. The International Institute for Strategic Studies.

From Surveillance Monocultures to Agroecological Defense: A Sovereignty-Centered Framework for Agricultural Cyberbiosecurity

Rolando Perez[1], Xavier Palmer[2], Lucas Potter[2], Dodzi Koku Hattoh[3], Salomey Afua Addo[4], and Srdjan Lesaja[5]

[1]Blue Marble Space Institute of Science, Seattle, WA, USA
[2]BiosView Labs, Dayton, Ohio, USA
[3]University of Ghana, Accra / Bonn Sustainable AI Lab, Bonn, Germany
[4]University of Cambridge, UK

rolando.perez@bmsis.org

Abstract: Agricultural cyberbiosecurity (CBS) scholarship may have overwhelmingly adopted centralized, national-security-oriented frameworks that leave community-governed and agroecologically-grounded alternatives unexplored. Recent AI safety research casts doubt on the viability of general-purpose, centralized safety regimes versus bounded or decentralized ones (Panigrahy and Sharan, 2025). Some leading proponents of centralized CBS paradigms propose surveillance architectures spanning unified biological intelligence (BIOINT), national bioaudit systems, and coordinated governance bodies that may cost billions of dollars. Centralized approaches often compellingly diagnose monoculture vulnerabilities, biosurveillance urgency, and the need for novel governance mechanisms. However, in these centralized architectures, communities receive biosecurity services but do not co-govern biosurveillance priorities, data use, or safety specifications, and agriculture remains a critical infrastructure to be defended from above. Even when such centralized paradigms propose distributed biological sensing, such as engineered sentinel plants and living biosensors, these systems require routine community-level care, maintenance, and trust relationships that centralized architectures cannot deliver, and will demand iterative updating as threat landscapes evolve. The empirical record confirms the costs of centralization: millions in ransom paid by JBS and a considerable percentage of U.S. grain production disrupted by a single cooperative's software failure (Yazdinejad et al., 2021; Cartwright and Cartwright, 2023). Coordination gaps are real, and new institutions are necessary to address them. However, coordination without corresponding community-level governance authority reproduces the monoculture pattern at the institutional level. This paper argues that agricultural defense must be rooted in the slowest and most durable layers of change identified by Stewart Brand's pace-layer framework: nature and culture. We advance three interlocking components: (1) agroecology as defensive architecture, in which biological diversity and functional redundancy constitute the primary cyberbiosecurity strategy; (2) sovereignty-preserving biosurveillance, in which community-governed federated sensing networks retain local data authority while enabling collective threat detection through privacy-preserving mechanisms; and (3) cooperative assurance from below, in which safety specifications are collectively deliberated and verified through participatory guarantee systems rather than centralized certification hierarchies (Dalrymple et al., 2024; Carroll et al., 2020; Manoj et al., 2025). This framework does not preclude centralized pathogen detection when necessary, but insists that top-level and community-level institutions be in relation with one another so that coordination flows bidirectionally rather than exclusively from above. Rather than claiming comprehensive safety guarantees, we demonstrate how bounded formal assurances, when integrated with agroecological resilience and community governance, can materially improve the security of food systems under real-world constraints.

Keywords: Biosecurity; Biofutures; Stewardship, Sovereignty, AI, Indigenous Data Governance

1. Introduction: Two Architectures for Agricultural Defense

As agriculture integrates artificial intelligence, networked sensors, and advanced biotechnologies, new risks emerge alongside potential benefits. AI models that optimize drone irrigation can be manipulated through data poisoning to trigger drought stress; botnets of compromised agritech devices could incapacitate farm networks during critical harvest windows (Aldhyani and Alkahtani, 2023; Pasca et al., 2025). Such disruptions cascade across food supply chains, amplify biosecurity incidents, and erode trust in markets and governance institutions (Duncan et al., 2019; Drape et al., 2021).

The cyberbiosecurity (CBS) discipline emerged to address convergent risks at the interface of biological systems, digital infrastructure, and national security. Foundational work by Murch et al. (2018) framed CBS around protecting the bioeconomy and mitigating national security threats, a paradigm extended by George (2019) through nuclear-deterrence analogies and reinforced in subsequent scholarship (Duncan et al., 2019; Stephen et al., 2023; Greenbaum, 2023). Recent CBS-adjacent policy-oriented contributions continue this orientation, advancing proposals for integrated biological intelligence systems, bioaudit mechanisms modeled on financial oversight, and expanded federal coordination structures (National Academies of Sciences, Engineering, and Medicine, 2018; Knight and Sureka, 2024; Endy et al., 2025). These proposals typically frame biology as an open, distributed network while advocating centralized architectures for biosurveillance, verification, and governance. At the same time, the literature makes substantive contributions, including articulating novel verification

mechanisms, emphasizing monoculture vulnerability, and situating biological security within long-term global food and supply-chain dependencies.

We support the view that coordination gaps are real. In the United States, fragmented biosecurity authority across multiple agencies may have historically hindered threat detection and response, and a new federal coordinating capacity is necessary. Centralised coordination that lacks corresponding community-level governance authority risks reproducing monoculture dynamics at the institutional level, consolidating surveillance, certification, and decision-making within a single hierarchical framework rather than distributing oversight across diverse local actors (Jannah et al., 2025; Powell et al., 2021). This paper argues that the top-level institutions needed must be in relation with community-level governance institutions, cooperative data governance bodies, participatory assurance networks, and farmer-led biosurveillance, so that coordination flows bidirectionally rather than exclusively from above.

Security is not synonymous with safety. Highly surveilled agricultural systems may be secured against external threats while remaining unsafe for farmers and ecosystems due to loss of autonomy, repair rights, and democratic oversight (Drape et al., 2021). The remainder of this paper examines the centralization paradox through the empirical cyberattack record (Section 2), establishes theoretical bounds on AI safety guarantees (Section 3), presents agroecology as defensive architecture (Section 4), develops a sovereignty-centered governance alternative (Section 5), models competing threat prioritizations (Section 6), and concludes with implementation pathways and limitations (Section 7).

2. A Centralization Paradox in CBS

Cyberbiosecurity analyses characterise agricultural cybersecurity as a collective-action problem shaped by fragmented ownership structures and uneven capacity, resulting in underinvestment in system-wide protection across interconnected supply chains (Duncan et al., 2019; Drape et al., 2021). Duncan et al. (2019) contend that agricultural cybersecurity suffers from a collective-action problem, as firms lack incentives to secure shared networks, and that the sector's diversity limits the effectiveness of centralized governance models (Duncan et al., 2019). Drape et al. (2021) report from a ~170-participant workshop that no group represented in the discussions had access to cybersecurity training or resources, highlighting widespread gaps in workforce preparedness and awareness around cyberbiosecurity across sectors, rather than evidence of training penetration at the farm level specifically. Greenbaum's (2023) comprehensive volume includes a chapter on safeguarding against social injustices, yet the CBS canon by and large remains institutional, bioeconomy-oriented, and state-centric. Across this literature, no work appeared to systematically address farmer data sovereignty, agroecological systems, community self-governance for security, or the risk that CBS frameworks themselves become instruments of surveillance and corporate consolidation.

Emerging biosecurity frameworks describe a distributed biological sensing network integrating routine antibody repertoire monitoring, bioengineered environmental sentinels, and real-time genomic sequencing, framing continuous bio-detection capacity as essential to national resilience against rapidly evolving or covert biological threats (Knight and Sureka, 2024; Endy et al., 2025). Concentrating detection, analysis, and response authority within a unified system reduces institutional diversity and increases systemic coupling, a structural condition that some cyberbiosecurity analyses identify as heightening vulnerability in tightly integrated agricultural and food system infrastructures (Peccoud et al., 2018; Murch et al., 2018). The proposed response, BIOINT, is to scale the same architectural logic rather than question it. BIOINT exemplifies the centralization paradox at its most ambitious scale. The proposed BIOINT architecture describes continuous, multi-domain metagenomic surveillance across environmental, agricultural, and human systems, coupled with cross-agency data integration and AI-enabled analytics to accelerate detection and attribution of emerging biological threats (Knight and Sureka, 2024; Endy et al., 2025).

The empirical record of agricultural cyberattacks reinforces the case that centralization can be a primary vulnerability. Ransomware activity in the food and agriculture sector demonstrates how attackers exploit centralized digital platforms and time-critical operations to generate disproportionate downstream effects across consolidated supply chains, preferentially timing intrusions to periods of heightened operational pressure such as planting, harvest, or peak processing windows (Cartwright and Cartwright, 2023; Yazdinejad et al., 2021). These risks are amplified by structural consolidation, which creates single points of failure whose compromise can affect substantial portions of production simultaneously (Perrow, 1999; Boyson, 2014; Ashley et al., 2022). The vulnerability profile of network-connected agricultural equipment makes this concrete. Across DEF CON 29 and DEF CON 30, security researcher Sick Codes demonstrated how weak defaults, exposed administrative interfaces, outdated operating systems, and insecure remote-management features in John Deere and related

platforms could enable high-privilege access to centralized operations systems and, by extension, large fleets of connected equipment (Sick Codes, 2021; Sick Codes, 2022). Complementary assessments identified more than 100 distinct vulnerabilities in Deere's corporate networks during a 10-day white-hat engagement (Manglicmot, 2022). Cyberbiosecurity scholarship has noted that software-locked agricultural machinery concentrates control over repair, modification, and operation within OEM ecosystems, and proprietary diagnostic restrictions further exacerbate exposure by preventing equipment owners from detecting or remediating vulnerabilities independently, thereby entrenching both risk and dependency (Wiens, 2018; Gambino, 2023; Stephen et al, 2023).

Cyberbiosecurity scholarship increasingly treats large-scale monoculture dominance, concentrated animal feeding operations, and globally interdependent food trade networks as structural amplifiers of systemic vulnerability, as agricultural uniformity reduces ecological resilience, intensification increases zoonotic emergence risk, and trade coupling enables cascading disruption (Altieri, 2009; Jones et al., 2013; Puma et al., 2015; Stephen et al, 2023). Yet some CBS experts and adjacent practitioners treat these exclusively as infrastructure problems requiring top-down monitoring and federal intervention. Farmers appear primarily as reluctant subjects of biosurveillance, never as active biosecurity agents whose knowledge, practices, and governance capacity constitute a defensive resource. The word "sovereignty" appears insufficiently in the CBS literature in an agricultural context. A recent biotechnology governance proposal utilises Society Readiness Levels to assess societal preparedness alongside technological maturity, noting that readiness varies across stakeholder groups and may be more decisive than technical development alone (Endy et al., 2025). This framework path offers a valuable lens that could be applied more broadly, including to CBS governance proposals themselves. For instance, the societal readiness of persistent metagenomic surveillance among the agricultural and Indigenous communities these systems would engage is likely low. Here, the SRL framework itself suggests a design insight worth further development: when societal readiness is low, and architecture depends on community participation, noncompliance becomes a systemic biosecurity vulnerability, not a problem of public education but a structural signal warranting alternative design approaches. Cyberbiosecurity is a necessary but insufficient component of agricultural defense. The centralization paradox is not that centralized capacity is unnecessary; it is that centralization unaccompanied by distributed capacity and governance authority converts a defensive architecture into a single point of failure. The following section examines why the AI safety literature reinforces this conclusion.

3. Bounded Assurance: What Safety Can and Cannot Guarantee

Panigrahy and Sharan (2025) prove an impossibility result showing that no AI system can simultaneously satisfy strong notions of safety, trust, and full generality. Drawing an analogy to Gödel's incompleteness theorems, they argue that safety- and trust-constrained systems must exclude some task instances that fall within the scope of general human problem-solving. The authors explicitly note that this result motivates the use of weaker, practical interpretations of safety and trust in real-world deployments.

Tegmark and Omohundro (2023) propose one of the most ambitious centralized safety architectures in the AI safety literature, advocating a hierarchy of provably compliant systems spanning software, hardware, and contractual enforcement mechanisms. Yet their concrete examples of tractable specifications, such as restricting DNA synthesizers from producing certain pathogens or imposing geofencing and time limits on AI hardware, appear to remain narrowly bounded rather than universal. Zero-knowledge proofs for DNA synthesis screening allow researchers to cryptographically verify that a requested sequence does not appear on a prohibited list without revealing the sequence itself, a genuinely useful application of privacy-preserving verification to a discrete and formally specifiable task. However, bounded systems address the verification problem without addressing the governance problem: who defines which sequences are dangerous, who maintains the reference database, and who decides the detection thresholds. Privacy-preserving verification without governance over the standards being verified is technical sophistication in the service of unexamined authority. Moreover, practitioners seem to confine this privacy-preserving logic to a single commercial checkpoint while deploying AI throughout BIOINT's analytical layer as an unconstrained force multiplier, without engaging whether this introduces the systemic vulnerability precisely of the kind Panigrahy and Sharan identify. Cryptographic privacy is offered to those with intellectual property to protect; pervasive data collection with "anonymized" and "opt-in" mentioned only in passing is proposed for everyone else.

Dalrymple et al. (2024), representing a broad coalition of AI safety researchers, articulate the Guaranteed Safe AI paradigm through three components: a world model, a safety specification, and a verifier. Their framework recognizes a spectrum of assurance mechanisms from formal proofs to probabilistic guarantees and empirical testing. Crucially, they emphasize that safety specifications must encode societal risk criteria and should ideally

be informed by collective deliberation, not solely as a technical problem but as a governance challenge. Their proposal for compositional verification, in which complex systems are decomposed into independently verified modules, points toward a distributed architecture of assurance. In agricultural cyberbiosecurity, different components may warrant distinct assurance regimes, ranging from formal verification in bounded control systems to empirical validation in complex agroecological monitoring.

Tomašev et al. (2025) advance a more radical departure by proposing distributional AGI safety, in which general intelligence emerges from the coordination of many specialized agents rather than a single monolithic system. Their framework emphasizes institutional and incentive-based controls, including reputation systems, cryptographic identity, and anti-monopoly safeguards, rather than universal internal alignment. This perspective maps naturally onto agricultural technology ecosystems, where safety and resilience emerge from interactions among heterogeneous tools, actors, and governance structures that eliminate reliance on centralized verification alone.

Collectively, these four contributions establish that complete safety is impossible (Panigrahy and Sharan, 2025), ambitious but practically limited to bounded applications (Tegmark and Omohundro, 2023), best understood as a spectrum requiring collective deliberation (Dalrymple et al., 2024), and most effectively achieved through distributed governance rather than monolithic control (Tomašev et al., 2025). Bounded examples from leading CBS paradigms, DNA synthesis screening, and geofenced hardware point toward this conclusion even as their overall architecture moves in the opposite direction. Agricultural AI safety requires bounded, task-specific, formally constrained subsystems, governed through collective deliberation, rather than comprehensive guarantees imposed over socio-ecological complexity.

4. Agroecology as Defensive Architecture

Agroecological systems, characterized by biodiversity, soil health, decentralization, and local knowledge, reduce biological, ecological, and cyber-biological vulnerabilities by design (Altieri et al.,2024, 2015; Perfecto et al., 2019). They limit monoculture-driven pathogen amplification, reduce reliance on centralized inputs, and constrain the scale at which failures can propagate. Critically, agroecology provides direct defense against biological threats, both natural and engineered. Polyculture cropping systems disrupt the host density and spatial continuity that pathogens require for epidemic transmission, making it structurally more difficult for any single pathogen to cause systemic failure across a diversified landscape. Recent research demonstrates that polycultures promote complex root exudate chemistry that recruits plant-beneficial microbes, some of which enhance plants' innate immune system, providing a form of endogenous biological defense that monoculture systems lack (Altieri et al., 2024). From a CBS perspective, agroecology narrows both biological and digital attack surfaces: diverse cropping systems, localized production practices, and reduced dependence on uniform digital platforms make coordinated large-scale disruption more difficult (IPES-Food, 2016).

The empirical evidence for agroecological resilience is substantial. Diversified farming systems showed greater resistance and faster recovery than monocultures following both Hurricane Mitch and Hurricane Maria (Altieri et al.,2024, 2015; Perfecto et al., 2019). The mechanism is functional redundancy: communities with more species performing similar functions are buffered against the loss of any single species (Biggs et al., 2020). As Ulanowicz (2018) established, "biodiversity, functional redundancy, and system persistence are all entwined."

These parallels between agroecological resilience and cybersecurity resilience map onto specific defensive functions instead of being merely metaphorical. Applying cues from Saltzer et al. (1984) and Littlewood and Strignini (2004) across biology and into security, functional redundancy corresponds to heterogeneous, independent security pathways: farms operating diverse sensor networks, communication protocols, and software platforms, such that the compromise of any single system does not cascade across the entire operation. Polyculture diversity corresponds to a multi-vendor, multi-protocol digital infrastructure that limits correlated failure modes. Biological pest control through predator-prey dynamics and beneficial microbial communities corresponds to community-operated intrusion detection, in which distributed human and technical monitoring identifies anomalies through local knowledge rather than centralized algorithmic surveillance. Plant immune promotion through soil microbiome health aligns with security-by-design principles, in which systems are hardened at the foundational level rather than defended through layered external monitoring. In each case, the agroecological principle provides resilience through diversity, distribution, and endogenous defense. At the same time, BIOINT's unified surveillance architecture importantly enables faster detection but concentrates it in a single system whose compromise or deception could simultaneously blind the entire network. Security engineering research independently confirms that heterogeneous designs reduce the impact of successful attacks by preventing compromise from propagating uniformly (Saltzer et al., 1984; Littlewood and Strigini,

2004; Kott et al., 2015), as the July 2024 CrowdStrike incident, the technological equivalent of a pest epidemic in a monoculture, dramatically illustrated (Alsowaigh, 2025).

The IPES-Food (2016) report identified structural lock-ins preventing transition from industrial to diversified agriculture, each with a precise agricultural technology analogue. Path dependency maps onto proprietary platform dependency; power concentration maps onto the six-corporation consolidation of more than 60% of global seed sales and the digital platforms farmers depend on (IPES-Food, 2016; Sauvagerd et al., 2024). The centralized paradigm architecture adds a governance lock-in layer on top of the existing corporate one.

Stewart Brand's pace-layer framework acknowledges that nature is the slowest-changing layer of civilization (Brand, 2018). The implication: biosecurity strategies rooted in ecological integrity and cultural practices, the slow layers, are structurally more durable than solutions operating at the governance and commerce layers, where most CBS proposals concentrate. Agroecological diversity, soil microbiome health, and the plant immune functions they support operate at the pace of nature and culture. The FAO's Ten Elements of Agroecology, endorsed by 197 member states, provides a complementary framework emphasizing diversity, resilience, and responsible governance (FAO, 2018; Barrios et al., 2019). The centralized CBS paradigm identifies the right analytical framework but proposes fast-layer solutions to slow-layer problems.

5. Sovereignty-Centered Governance: From Bioaudits to Cooperative Assurance

Endy et al (2025)'s bioaudit proposal is among the report's most unique contributions. Modeled on financial auditing, it envisions independent third-party evaluations of biosafety and biosecurity practices, a credentialing process analogous to public accountant certification, and a Bioaudit Oversight Board modeled on the Public Company Accounting Oversight Board. The tiered verification framework, in which nations choose between unverified, self-verified, or multilaterally verified status, offers a pragmatic pathway for BWC modernization. The financial auditing analogy reveals structural risks that warrant further development. The history of financial auditing demonstrates that third-party verification within centralized hierarchies is susceptible to systemic capture: Enron, the 2008 financial crisis, and Wirecard all occurred under regimes of ostensibly independent audit oversight. A bioaudit board operating within the same institutional logic could benefit from mechanisms to mitigate analogous risks, particularly by clarifying who defines the standards against which audits are conducted.

Contemporary global bio-surveillance proposals describe large-scale biological data collection infrastructures incorporating metagenomic environmental sequencing, adaptive immune repertoire monitoring, and engineered living biosensors, often invoking anonymization and opt-in participation as governance safeguards that remain under-specified in technical detail; these are important functions via BIONT (Georgiou et al., 2014; Crits-Christoph et al., 2021; Mittelstadt and Floridi, 2016). This centralized CBS paradigm literature so far appears to not have deeply engaged with Indigenous data sovereignty frameworks, community consent mechanisms, or the political economy of biological surveillance data; data governance has considerable room to grow. This omission is particularly striking given that the world's largest technical professional organization has now codified Indigenous data governance as a technical standard: IEEE 2890-2025, the Recommended Practice for Provenance of Indigenous Peoples' Data, establishes parameters for recording data provenance related to Indigenous Peoples, their cultures, lands, and knowledge systems, and defines "data actors" to include non-human entities such as devices, applications, and AI systems (IEEE, 2025). The CARE Principles for Indigenous Data Governance, foregrounding collective benefit, authority to control, responsibility, and ethics, provide the normative foundation that IEEE 2890-2025 operationalizes (Carroll et al., 2023; IEEE, 2025). Leonelli's (2024) ethnographic research at Ghana's Crops Research Institute revealed that the push to digitize crop data as "global commons" can function as "a sophisticated form of bioprospecting and surveillance, whereby local knowledge of crops and their uses is harnessed and mined in ways that damage the very communities that produced it." Her subsequent work argues that dominant AI is underpinned by an anti-human philosophy of mastering nature, proposing instead "Environmental Intelligence" as the ability to live in and with nature (Leonelli, 2025), a vision aligned with the agroecological framework advanced here.

The alternative is cooperative assurance from below. Smallholder-led cooperatives could form associations to perform justice-accountable audits, retain shared access keys to models and datasets, and implement consent-by-design governance processes. Without such mechanisms, federated learning reproduces centralization in model control and monetization even as it decentralizes data storage: so long as smallholders lack meaningful governance over models and control over outputs, they remain data laborers rather than data owners (Zalik and Zalik, 2023). Manoj et al. (2025) demonstrate through AgriFLChain that blockchain-assisted decentralized identifiers and smart contracts can authenticate participants and govern model updates without reliance on a

central authority. The critical governance questions are: who decides which models are trained, who controls aggregation, and who sets the privacy parameters. In a sovereignty-centered framework, cooperative data governance boards, modeled on long-standing producer-led institutions such as the National Cooperative Dairy Herd Improvement Program (Wiggans et al., 2011), set research priorities through collective deliberation. Communities determine the differential privacy parameters, deciding how much privacy to trade for model accuracy.

Often, zero-knowledge proof proposals apply privacy-preserving verification to one narrow commercial checkpoint. Yet ZKPs, federated learning, and differential privacy belong to the same family of cryptographic approaches to decentralized trust, all solving structurally similar problems: deriving collective intelligence or verification from distributed data without requiring any party to surrender data to a central authority. Centralized CBS paradigms often glimpse this principle in one application but cannot follow through because generalizing it would require distributing governance authority to the communities whose data flows through the system, contradicting the centralized architecture on which the rest of the systems depend. Federated learning with community-governed differential privacy completes what the ZKP proposal begins, extending privacy-preserving verification from a single commercial transaction to the entire agricultural data ecosystem. The technical and institutional components for this alternative already exist. OpenTEAM and its farmOS platform provide interoperable, farmer-driven digital infrastructure through open APIs; L'Atelier Paysan demonstrates farmer-led technological sovereignty through open-source agricultural machinery (IPES-Food, 2016); and participatory guarantee systems in contexts from Mexico to India demonstrate that peer verification with social accountability can function as credible alternatives to centralized certification (Nelson et al., 2016; Ruder and Wittman, 2025).

6. Threat Model Considerations: Competing Risk Prioritizations and Discussions

Any framework for agricultural defense must specify what it defends against, at what cost, and what it sacrifices. The sovereignty-centered framework advanced here, and the centralized architecture proposed by the aforementioned dominant CBS paradigms, represent competing risk prioritizations rather than an ideological disagreement.

For ransomware and supply-chain attacks, decentralised or sovereignty-centred architectures offer structural advantages. Highly centralised systems create single points of failure, and in tightly coupled supply networks such failures can propagate as cascading disruptions across interdependent infrastructures (Rinaldi et al., 2001; Scheibe and Blackhurst, 2018; Ivanov et al., 2019). Decentralized systems operating on heterogeneous platforms with community-governed data infrastructure eliminate the high-value targets that ransomware operators preferentially seek. The centralized CBS architecture, which consolidates biosurveillance, certification, and coordination within unified federal systems, replicates the vulnerability pattern that enabled the JBS and cooperative software incidents at the governance level.

For novel engineered pathogen detection, the centralized framework holds an advantage in aggregation speed. Distributed sentinel systems may detect more slowly than a unified BIOINT system capable of cross-referencing metagenomic data across regions in real time. This paper does not claim that community governance eliminates the need for centralized analytical capacity in this threat class. However, three qualifications temper this concession. First, centralized systems may have failed to detect SARS-CoV-2 for approximately two months despite existing biosurveillance infrastructure (Pekar et al., 2021). Second, distributed systems are harder to spoof or systematically blind precisely because there is no single point of deception; an adversary must compromise multiple independent detection networks simultaneously rather than one unified system. Third, community-level monitoring creates human intelligence layers, farmers noticing anomalies in their crops, livestock, or soil that purely technical surveillance misses and that cannot be replicated by remote metagenomic sequencing. The agroecological foundation described in Section 4 adds a further layer: polyculture systems that disrupt pathogen transmission and promote plant immune function provide passive biological defense, reducing the probability that an engineered pathogen achieves epidemic spread in the first place (Altieri et al., 2024).

The sovereignty-centered framework addresses a third threat class that centralized CBS paradigms do not sufficiently appear to recognize: corporate data extraction and surveillance capitalism. BIOINT is a massive biological data-collection infrastructure that lacks engagement with the political economy of who controls that data or profits from it. Agricultural data is already subject to oligopolistic platformisation, in which four corporations control the majority of the digital platforms farmers depend on (IPES-Food, 2016; Sauvagerd et al., 2024). The FTC's 2025 suit against John Deere for unlawful repair cost inflation and for documenting $3 billion in annual farmer losses from tractor downtime illustrates how proprietary control functions as an ongoing

extraction mechanism (Rimmer, 2025). When farmers buy older equipment to avoid software on modern machinery, the cybersecurity implications are clear: proprietary systems actively prevent the distributed defense capacity that biosecurity requires (Wiens, 2018). A framework that treats corporate consolidation as external to biosecurity cannot address this threat class. A sovereignty-centered framework treats it as foundational.

Centralized CBS paradigms optimize for threat class 2 at the expense of threat classes 1 and 3. The sovereignty-centered framework optimizes differently, accepting a bounded trade-off in centralized detection speed in exchange for structural resilience against supply-chain attacks, systemic resistance to data extraction, and the endogenous biological defense that agroecological diversity provides against pathogens of any origin.

7. Conclusion: Toward Implementation

CBS needs a paradigm that treats centralization as a threat class, agroecology as defensive architecture, and sovereignty as a security requirement. This paper has argued that the strongest expression of the centralized paradigm, BIOINT, diagnoses the right threats but proposes an architecture that replicates the monoculture vulnerability identified in agriculture. The alternative advanced here, built on agroecological defense, sovereignty-preserving biosurveillance, and cooperative assurance from below, addresses threat classes that centralized frameworks structurally cannot see while engaging those they prioritize from a more resilient foundation.

Implementation can proceed through existing institutional infrastructure. In an initial phase, pilot deployments would build on operational cooperative platforms already in use: OpenTEAM and farmOS for interoperable farm data management, the National Cooperative Dairy Herd Improvement Program model for producer-led data governance, and participatory guarantee systems for peer-verified compliance (Nelson et al., 2016; Wiggans et al., 2011). A second phase would integrate decentralized federated learning with cooperative governance boards, using AgriFLChain-style blockchain-assisted authentication and community-set differential privacy parameters to enable collective threat detection without centralized data aggregation (Manoj et al., 2025). A third phase would establish a cross-cooperative federation, connecting regional sovereignty-preserving biosurveillance networks into broader detection systems that retain local governance authority over data and safety specifications. This phased pathway builds community governance capacity from below, with the two architectures designed to integrate bidirectionally rather than one subordinating the other.

This paper does not claim that technical safeguards can eliminate risk, nor does it claim that community governance is inherently immune to failure. Cooperative structures can fail through exclusion, elite capture, uneven capacity, or internal conflict. The sovereignty-centered framework does not eliminate the need for some centralized capacity for pathogen detection and genomic analysis, particularly for novel engineered threats. Agroecological transition is itself constrained by the very lock-ins this paper documents: path dependency, corporate consolidation, and policy environments that continue to privilege monoculture production (IPES-Food, 2016).

Stewart Brand's Pace Layer Framework notes that nature is the slowest-changing layer of civilization (Brand, 2018). Agricultural defense built on ecological integrity, soil microbiome health, community knowledge, and cooperative governance and slow layers provides a lasting and democratic foundation for any federal coordinating body, bioaudit regime, or AI-driven surveillance system. This paper argues that durable biosecurity emerges when the speed of detection is in relation with the depth of the systems, ecological and social, in which defense is rooted.

Acknowledgements

Special thanks go to the Spirit of Asilomar Next Generation Leaders Cohort and Indigenous Stewards (and their supporters). We acknowledge Dr. Jessica Snyder for proofreading and helpful comments.

Ethics Declaration

Ethical Clearance was not required.

AI Declaration

AI tools in Grammarly, Claude, ChatGPT, and Google Docs were used for literature searches and reviews, cross-referencing, formatting, spell-checking, and overall paper organization of pre-written text.

References

Aldhyani, T.H.H. and Alkahtani, H. (2023) 'Cyber security for detecting distributed denial of service attacks in Agriculture 4.0: Deep learning model', *Mathematics*, 11(1), p. 233. Available at: https://doi.org/10.3390/math11010233.

Alsowaigh, R.E. (2025) 'Crowdstrike causes global Microsoft outage: a case study', Journal of Information Security and Cybercrimes Research, 8(1), pp. 63–76. https://doi.org/10.26735/QHDD4798

Altieri, M.A., 2009. Agroecology, small farms, and food sovereignty. Monthly review, 61(3), pp.102-113.

Altieri, M.A., Nicholls, C.I., Henao, A. et al. (2015) 'Agroecology and the design of climate change-resilient farming systems', Agronomy for Sustainable Development, 35, pp. 869–890. doi:10.1007/s13593-015-0285-2.

Altieri, M.A., Nicholls, C.I., Dinelli, G. and Negri, L. (2024) 'Towards an agroecological approach to crop health: reducing pest incidence through synergies between plant diversity and soil microbial ecology', *npj Sustainable Agriculture*, 2, p. 6. Available at: https://doi.org/10.1038/s44264-024-00016-2.

Ashley, T., Gourisetti, S.N.G., Brown, N. and Bonebrake, C. (2022) 'Aggregate attack surface management for network discovery of operational technology', *Computers & Security*, 123, p. 102939. Available at: https://doi.org/10.1016/j.cose.2022.102939.

Barrios, E. *et al.* (2020) 'The 10 Elements of Agroecology: enabling transitions towards sustainable agriculture and food systems through visual narratives', *Ecosystems and People*, 16(1), pp. 230–247. doi: 10.1080/26395916.2020.1808705.

Biggs, C.R., Yeager, L.A., Bolser, D.G., Bonsell, C., Dichiera, A.M., Hou, Z., Keyser, S.R., Khursigara, A.J., Lu, K., Muth, A.F., Negrete, B. Jr. and Erisman, B.E. (2020) 'Does functional redundancy affect ecological stability and resilience? A review and meta-analysis', *Ecosphere*, 11(7), p. e03184. Available at: https://doi.org/10.1002/ecs2.3184.

Boyson, S. (2014) 'Cyber supply chain risk management: Revolutionizing the strategic control of critical IT systems', *Technovation*, 34(7), pp. 342–353. Available at: https://doi.org/10.1016/j.technovation.2014.02.001.

Brand, S. (2018) 'Pace Layering: How Complex Systems Learn and Keep Learning', *Journal of Design and Science* [Preprint]. doi:10.21428/7f2e5f08.

Carroll, S.R. et al. (2020) 'The CARE Principles for Indigenous Data Governance', *Data Science Journal*, 19, p. 43, pp. 1–12. Available at: https://doi.org/10.5334/dsj-2020-043.

Cartwright, A. and Cartwright, E. (2023) 'The economics of ransomware attacks on integrated supply chain networks', *Digital Threats*, 4(4), Article 56, pp. 1–14. Available at: https://doi.org/10.1145/3579647.

Crits-Christoph, A., Kantor, R.S., Olm, M.R., Whitney, O.N., Al-Shayeb, B., Lou, Y.C., Flamholz, A., Kennedy, L.C., Greenwald, H., Hinkle, A. and Hetzel, J., 2021. Genome sequencing of sewage detects regionally prevalent SARS-CoV-2 variants. MBio, 12(1), pp.10-1128. doi: 10.1128/mbio.02703-20

Dalrymple, D., Skalse, J., Bengio, Y., Russell, S., Tegmark, M., Seshia, S., Omohundro, S., Szegedy, C., Goldhaber, B., Ammann, N., Abate, A., Halpern, J., Barrett, C., Zhao, D., Zhi-Xuan, T., Wing, J. and Tenenbaum, J. (2024) 'Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems', *arXiv preprint*, arXiv:2405.06624v2. Available at: https://arxiv.org/abs/2405.06624 (Accessed: 15 February 2026), https://doi.org/10.48550/arXiv.2405.06624.

Drape, T., Magerkorth, N., Sen, A., Simpson, J., Seibel, M., Murch, R.S. and Duncan, S.E. (2021) 'Assessing the Role of Cyberbiosecurity in Agriculture: A Case Study', *Frontiers in Bioengineering and Biotechnology*, 9, 737927. doi:10.3389/fbioe.2021.737927.

Duncan, S.E., Reinhard, R., Williams, R.C., Ramsey, F., Thomason, W., Lee, K., Dudek, N., Mostaghimi, S., Colbert, E. and Murch, R. (2019) 'Cyberbiosecurity: A New Perspective on Protecting U.S. Food and Agricultural System', *Frontiers in Bioengineering and Biotechnology*, 7, 63. doi:10.3389/fbioe.2019.00063.

Endy, D., Moront, S., Alexopoulos, V.A., Patel, R., Jain, R. and Bennett, B. (October 2025) *Biosecurity Really: A Strategy for Victory*. Stanford, CA: Hoover Institution, Bio-Strategies & Leadership Initiative.

FAO, 2018. The 10 Elements of Agroecology: Guiding the Transition to Sustainable Food and Agricultural Systems. Rome: Food and Agriculture Organization of the United Nations. Available at: https://openknowledge.fao.org/server/api/core/bitstreams/3d7778b3-8fba-4a32-8d13-f21dd5ef31cf/content (Accessed: 15 February 2026).

Gambino, A.J. (2023) 'Right to Repair: Whose Right is it Anyway?', *Tennessee Journal of Business Law*, 25(1), pp. 1–30. doi:10.70658/4486-1457.1650.

George, A.M. (2019) 'The National Security Implications of Cyberbiosecurity', *Frontiers in Bioengineering and Biotechnology*, 7, 51. doi:10.3389/fbioe.2019.00051.

Georgiou, G., Ippolito, G., Beausang, J. et al. (2014) 'The promise and challenge of high-throughput sequencing of the antibody repertoire', *Nature Biotechnology*, 32, pp. 158–168. doi:10.1038/nbt.2782.

Greenbaum, D. (2023) *Cyberbiosecurity : a new field to deal with emerging threats*. Cham, Switzerland : Springer. Available at: https://doi.org/10.1007/978-3-031-26034-6.

IEEE (2025) IEEE Recommended Practice for Provenance of Indigenous Peoples' Data. IEEE Std 2890-2025, pp. 1–25. Available at: https://ieeexplore.ieee.org/document/11302960 (Accessed: 15 February 2026).

IPES-Food (2016) *From Uniformity to Diversity: A Paradigm Shift from Industrial Agriculture to Diversified Agroecological Systems*. Brussels: International Panel of Experts on Sustainable Food Systems. Available at: http://www.ipes-food.org (Accessed: 15 February 2026).

Ivanov, D., Sokolov, B. and Dolgui, A. (2014) 'The Ripple effect in supply chains: trade-off 'efficiency-flexibility-resilience' in disruption management', *International Journal of Production Research*, 52(7), pp. 2154–2172. doi: 10.1080/00207543.2013.858836.

Jannah, R., Anjum, B.E., Lardner, C., Chappell, C., Palmer, X.-L., Perez, R., Mitra, A., Camenares, D., Seah, A., Kong, D., Elcock, L., Thaweechuen, J. and Flores, W.N. (2025) *6.3 Community Biology: Advancing Responsible Biotech Innovation*. Houston, TX: Rice University. doi:10.25611/3QHJ-W351.

Jones, B.A., Grace, D., Kock, R., Alonso, S., Rushton, J., Said, M.Y., McKeever, D., Mutua, F., Young, J., McDermott, J. and Pfeiffer, D.U. (2013) 'Zoonosis emergence linked to agricultural intensification and environmental change', *Proceedings of the National Academy of Sciences*, 110(21), pp. 8399–8404. doi:10.1073/pnas.1208059110.

Knight, T. and Sureka, S. (2024) 'A New Paradigm for Threat Agnostic Biodetection: Biological Intelligence (BIOINT)', *Health Security*, 22(1), pp. 31–38. doi:10.1089/hs.2023.0072.

Kott, A., Alberts, D.S. and Wang, C. (2015) 'Will Cybersecurity Dictate the Outcome of Future Wars?', *Computer*, 48(12), pp. 98–101. doi:10.1109/MC.2015.359.

Leonelli, S. (2024) 'Globalizing plant knowledge beyond bioprospecting?', *History of Anthropology Review*, 48. Available at: https://histanthro.org/notes/globalizing-plant-knowledge/ (Accessed: 15 February 2026).

Leonelli, S. (2025) 'Rejoinder: Let Us Rescue Intelligence from Anti-Human AI', *Harvard Data Science Review*, 7(4). doi:10.1162/99608f92.4a868f98.

Littlewood, B. and Strigini, L. (2004) 'Redundancy and Diversity in Security', in Samarati, P., Ryan, P., Gollmann, D. and Molva, R. (eds.) *Computer Security – ESORICS 2004*. Lecture Notes in Computer Science, vol. 3193. Berlin: Springer, pp. 423–438. doi:10.1007/978-3-540-30108-0_26.

Manglicmot, M. (2022) 'Hackers are coming for our Thanksgiving turkeys and John Deere tractors. It's time to reevaluate America's food security', *Fortune*, 23 November. Available at: https://fortune.com/2022/11/23/hackers-thanksgiving-turkeys-john-deere-tractors-food-security-ransomware-cyberattacks-tech-mark-manglicmot/ (Accessed: 15 February 2026).

Manoj, T., Makkithaya, K. and Narendra, V. (2025) 'A Blockchain-Assisted Trusted Federated Learning for Smart Agriculture', *SN Computer Science*, 6, 221. doi:10.1007/s42979-025-03672-4.

Mittelstadt, B.D. and Floridi, L. (2016) 'The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts', *Science and Engineering Ethics*, 22(2), pp. 303–341. doi:10.1007/s11948-015-9652-2.

Murch, R.S., So, W.K., Buchholz, W.G., Raman, S. and Peccoud, J. (2018) 'Cyberbiosecurity: An Emerging New Discipline to Help Safeguard the Bioeconomy', *Frontiers in Bioengineering and Biotechnology*, 6, 39. doi:10.3389/fbioe.2018.00039.

National Academies of Sciences, Engineering, and Medicine (2018) Biodefense in the Age of Synthetic Biology. Washington, DC: The National Academies Press. https://doi.org/10.17226/24890

Nelson, E.M., Gómez Tovar, L., Gueguen, E., Humphries, S., Landman, K. and Schwentesius de Rindermann, R. (2016) 'Participatory guarantee systems and the re-imagining of Mexico's organic sector', *Agriculture and Human Values*, 33, pp. 373–388. doi: 10.1007/s10460-015-9615-x

Panigrahy, R. and Sharan, V. (2025) 'Limitations on safe, trusted, artificial general intelligence', *arXiv preprint*, arXiv:2509.21654. doi:10.48550/arXiv.2509.21654.

Pasca, E.M., Delinschi, D., Erdei, R., Baraian, I. and Matei, O.D. (2025) 'A Vulnerable-by-Design IoT Sensor Framework for Cybersecurity in Smart Agriculture', *Agriculture*, 15(12), 1253. doi:10.3390/agriculture15121253.

Peccoud, J., Gallegos, J.E., Murch, R., Buchholz, W.G. and Raman, S., 2018. Cyberbiosecurity: from naive trust to risk awareness. Trends in biotechnology, 36(1), pp.4-7.doi:10.1016/j.tibtech.2017.10.012

Pekar, J. et al. (2021) 'Timing the SARS-CoV-2 index case in Hubei province', *Science*, 372, pp. 412–417. doi:10.1126/science.abf8003.

Perfecto, I., Hajian-Forooshani, Z., Iverson, A. et al. (2019) 'Response of Coffee Farms to Hurricane Maria: Resistance and Resilience from an Extreme Climatic Event', *Scientific Reports*, 9, 15668. doi:10.1038/s41598-019-51416-1.

Perrow, C. (1999) *Normal Accidents: Living with High-Risk Technologies*. Updated edn. Princeton, NJ: Princeton University Press, ISBN: 9780691004129.

Powell, E., Akogo, D., Potter, L. and Palmer, X.L., 2021, October. Co-leadership and cross-pollination of university and DIY bio spaces: an exploration in consideration of Biocybersecurity. In Proceedings of the Future Technologies Conference (pp. 610-621). Cham: Springer International Publishing.

Puma, M.J., Bose, S., Chon, S.Y. and Cook, B.I., 2015. Assessing the evolving fragility of the global food system. Environmental Research Letters, 10(2), p.024007.

Rimmer, M., 2025. Tractor rage: Intellectual property, agriculture, competition policy, and the right to repair. *IIC-International Review of Intellectual Property and Competition Law*, 56(1), pp.115-152. doi: 10.1007/s40319-024-01538-5

Rinaldi, S.M., Peerenboom, J.P. and Kelly, T.K., 2001. Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE control systems magazine, 21(6), pp.11-25. doi: 10.1109/37.969131

Ruder, S.L. and Wittman, H., 2025. Agricultural data governance from the ground up: Exploring data justice with agri-food movements. *Big Data and Society*, 12(1), p.20539517251330182. doi:10.1177/20539517251330182

Saltzer, J.H., Reed, D.P. and Clark, D.D., 1984. End-to-end arguments in system design. *ACM Transactions on Computer Systems*, 2(4), pp.277-288. doi:10.1145/357401.357402

Sauvagerd, M., Mayer, M. and Hartmann, M., 2024. Digital platforms in the agricultural sector: Dynamics of oligopolistic platformisation. *Big Data and Society*, 11(4), p.20539517241306365. doi:10.1177/20539517241306365

Scheibe, K.P. and Blackhurst, J., 2018. Supply chain disruption propagation: a systemic risk and normal accident theory perspective. International journal of production research, 56(1-2), pp.43-59. doi:10.1080/00207543.2017.1355123

Sick Codes, 2021. DEF CON 29: The agricultural data arms race: Exploiting a tractor load of vulns [online video]. YouTube. Available at: https://youtu.be/zpouLO-GXLo

Sick Codes, 2022. DEF CON 30: Hacking the farm: Breaking badly into agricultural devices [online video]. YouTube, 18 August. Available at: https://youtu.be/z2_TLz9TpwY

Stephen, S., Alexander, K., Potter, L. and Palmer, X.L., 2023. Implications of cyberbiosecurity in advanced agriculture. doi:10.34190/iccws.18.1.995

Tegmark, M. and Omohundro, S., 2023. Provably safe systems: the only path to controllable AGI. *arXiv preprint arXiv:2309.01933*.

Tomašev, N., Franklin, M., Jacobs, J., Krier, S. and Osindero, S., 2025. Distributional AGI safety. *arXiv preprint arXiv:2512.16856*.

Ulanowicz, R.E., 2018. Biodiversity, functional redundancy and system stability: subtle connections. *Journal of The Royal Society Interface*, 15(147), p.20180367. doi:10.1098/rsif.2018.0367

Wiens, K., 2018. We can't let John Deere destroy the very idea of ownership. Wired, 7 February. Available at: https://www.wired.com/story/john-deere-farmers-right-to-repair/

Wiggans, G.R., VanRaden, P.M. and Cooper, T.A., 2011. The genomic evaluation system in the United States: Past, present, future. *Journal of Dairy Science*, 94(6), pp.3202-3211. doi: 10.3168/jds.2010-3866

Yazdinejad, A., Zolfaghari, B., Azmoodeh, A., Dehghantanha, A., Karimipour, H., Fraser, E., Green, A.G., Russell, C. and Duncan, E., 2021. A review on security of smart farming and precision agriculture: Security aspects, attacks, threats and countermeasures. *Applied Sciences*, 11(16), p.7518. doi: /10.3390/app11167518

Zalik, K.R. and Zalik, M., 2023. A review of federated learning in agriculture. *Sensors*, 23(23), p.9566. doi: 10.3390/s23239566

From Hoax to Reality: Deepfake-driven Misinformation and the Death of Ozzy Osbourne

Alexander Pfeiffer[1], Nanditha Krishna[3], Thomas Wernbacher[2] and Walter Seböck[1]
[1]Department for Security Studies, University for Continuing Education Krems, Austria
[2]Department for Arts and Culture, University for Continuing Education Krems, Austria
[3] Independent Researcher

alexander.pfeiffer@donau-uni.ac.at
walter.seboeck@donau-uni.ac.at
Thomas.wernbacher@donau-uni.ac.at
Nandithakrish4@gmail.com

Abstract: Deepfakes are AI-generated images, videos, and texts that convincingly mimic real individuals. In recent years, such forgeries have proliferated across social media, and their role in fraud cases has increased markedly. However, detection tools often fail in real-world conditions; open-source detectors typically perform only half as well on "in-the-wild" content compared to curated test sets. This performance gap heightens the risk that fabricated content will undermine public trust and foster a climate of suspicion in which even authentic recordings are questioned—a phenomenon known as the "liar's dividend." In this case study, we examine how the July 2025 death of rock icon Ozzy Osbourne became a focal point for deepfake-driven misinformation and public speculation. Following a seated farewell concert in Birmingham, multiple synthetic videos surfaced, including one in which a digitally recreated Osbourne claimed he knew he was about to die. The clips sparked speculation about assisted suicide, prompting his daughter Kelly to publicly denounce the videos as fake and criticise those who shared them. When Osbourne died two weeks later, some commentators treated the deepfake as prophetic, fuelling conspiracy theories and amplifying public grief. This case illustrates broader ethical and governance challenges related to generative AI. Voice cloning and face-swapping services can create convincing media from minimal training data, yet developers rarely address issues of consent or privacy when sourcing material. Psychological factors such as fear of missing out (FOMO) encourage the viral spread of sensational content without verification. This paper's primary contribution is a theoretical synthesis, which integrates existing technical, psychological, and governance perspectives on deepfake-driven misinformation through a single illustrative case study. Effective countermeasures must combine technical innovations—such as blockchain-based provenance tracking and robust detection—with clear policy frameworks that regulate data use and require transparent labelling of synthetic media. Public education remains essential to help individuals recognise deepfakes and preserve trust in authentic digital communication.

Keywords: Deepfakes, Misinformation, Fake news, Liar's dividend, Digital trust, FOMO

1. Introduction

The proliferation of generative artificial intelligence (GenAI) has democratised the creation of synthetic media, enabling anyone with an internet connection to produce highly realistic yet fabricated images, audio, and video content, commonly known as deepfakes (Lundberg and Mozelius, 2025). Whilst this technology offers creative and educational affordances, its potential for malicious use has become a significant societal concern. Deepfakes are increasingly implicated in a range of harmful activities, including the spread of disinformation, character defamation, financial fraud, and the manipulation of democratic processes (Lundberg and Mozelius, 2025; Brennan Center for Justice, 2023; Pfeiffer and Krishna, 2025). The core challenge lies in the technology's ability to erode public trust in digital communication, creating an environment where the authenticity of any piece of media can be called into question. This erosion of trust is not merely a theoretical risk; it is a tangible threat that has been observed in numerous real-world incidents, where manipulated content has been used to mislead the public and sow discord. Recent advances in generative AI, particularly Google's Veo 3 model announced in May 2025, have dramatically escalated the sophistication and accessibility of synthetic media creation, enabling the generation of photorealistic videos with synchronised audio that are nearly indistinguishable from authentic content (Pfeiffer and Krishna, 2025). In its first week of release, users posted fake news segments in multiple languages, including fabricated death announcements and political news conferences. One commentator described Veo 3 as a potential "death knell for truth on the internet" (PCMag, 2025).

The Ozzy Osbourne case of July 2025 serves as a stark and poignant example of the real-world impact of deepfake technology. The incident, which unfolded in the weeks leading up to the rock icon's death, illustrates the complex interplay between synthetic media, public speculation, and the psychological vulnerabilities that drive the spread of misinformation. Multiple deepfake videos, appearing to show a frail Osbourne prophesying his own demise, were widely circulated on social media platforms, triggering a maelstrom of rumours and conspiracy theories (Amelinckx, 2025). The videos' apparent prescience, following Osbourne's actual death on

22 July 2025, lent them a veneer of credibility in the eyes of some, demonstrating how deepfakes can exploit and amplify public grief and confusion. The incident is documented in the OECD AI Incidents Monitor as "AI-Generated Deepfake Video Falsely Claims Ozzy Osbourne Is Dying" (incident date: 12 July 2025), classified as an AI Incident with reputational, psychological, and public interest harms (OECD AI Policy Observatory, 2025). This case study provides a valuable lens through which to examine the multifaceted challenges posed by deepfakes, from the technical limitations of detection tools to the ethical and governance vacuums in which this technology currently operates. By dissecting the Osbourne case, we can gain a deeper understanding of the mechanisms by which deepfakes exert their influence and explore potential countermeasures to mitigate their harmful effects. This paper's primary contribution is a theoretical synthesis that integrates existing technical, psychological, and governance perspectives on deepfake-driven misinformation. Rather than advancing a new empirical study or formal theory, we use the Ozzy Osbourne case as a single, illustrative example to put together these disparate threads of analysis. This approach allows for a broader understanding of the deepfake phenomenon, guiding how technological capabilities, human psychology, and regulatory frameworks interact in a real-world context

This paper will proceed as follows. First, it will provide a brief overview of deepfake technology and the current state of detection capabilities, highlighting the significant performance gap between academic benchmarks and real-world applications. Second, it will present a detailed analysis of the Ozzy Osbourne case study, tracing the timeline of events and examining the public's reaction to the deepfake videos. Third, it will look into the psychological factors, such as the 'fear of missing out' (FOMO), that contribute to the viral spread of sensationalist and unverified content. Finally, it will explore the broader ethical and governance challenges associated with generative AI, including the issues of consent in data training and the need for robust policy frameworks. Effective countermeasures must combine technical innovations such as blockchain-based provenance tracking with robust detection frameworks and clear labelling of synthetic media.

2. The Technological Challenge: Deepfakes and the Detection Arms Race

Deepfakes are the product of sophisticated machine learning techniques, specifically generative adversarial networks (GANs), which involve two neural networks—a generator and a discriminator—competing against each other to produce increasingly realistic synthetic media. The generator creates the fake content, whilst the discriminator attempts to distinguish it from authentic content. This adversarial process results in a rapid improvement in the quality and believability of deepfakes, making them increasingly difficult to detect with the naked eye. The accessibility of deepfake creation tools has also grown exponentially, with numerous online services and open-source software allowing users to create convincing forgeries with minimal technical expertise (Lundberg and Mozelius, 2025). This democratisation of deepfake technology, whilst having some positive applications in areas like entertainment and education, has also lowered the barrier to entry for malicious actors seeking to exploit it for nefarious purposes. The landscape has become dramatically more challenging with recent technological advances that can create clips that are "nearly indistinguishable from real ones," complete with dialogue, soundtracks, sound effects, and adherence to real-world physics (Pfeiffer and Krishna, 2025)[1].

The rapid advancement of deepfake generation has triggered a corresponding arms race in detection technology. Researchers and technology companies are developing a variety of methods to identify synthetic media, ranging from the analysis of digital artefacts and inconsistencies in the generated content to the use of blockchain-based provenance tracking to verify the origin and history of a piece of media (Content Authenticity Initiative, 2025; Tech Policy Press, 2025; Pfeiffer and Krishna, 2025). However, the effectiveness of these detection tools in real-world scenarios remains a significant concern. A recent study by Chandra et al. (2025), which introduced a new benchmark of in-the-wild deepfakes, found that the performance of open-source state-of-the-art detection models exhibited a marked performance degradation when evaluated on real-world content, with a decrease in the area under the curve (AUC) of approximately 50% for video models, 48% for audio models, and 45% for image models compared to their performance on academic datasets. This performance gap is a critical vulnerability, as it means that many of the deepfakes circulating on social media and other online platforms are likely to evade detection by automated systems. Even commercial detection models, whilst outperforming their open-source counterparts, still fall short of the accuracy achieved by human forensic analysts, highlighting the ongoing challenge of developing robust and reliable detection methods

[1]One of the authors can confirm this from personal experience. After a playdate, the mother of the other child sent a photo of both children to her father. He uploaded it to GROG without any hesitation and created an extremely lifelike video showing the two children dancing and falling over. This shows that the tools are easy to use even for computer-savvy people aged 80+, but that media literacy was completely lacking, at least in this case. In addition to the cyberbullying aspect, the overarching issue is, of course, the provision of training data from image material that does not belong to one person. This also applies to images of minors and even small children.

(Chandra et al., 2025). The study's findings underscore a troubling reality: existing approaches, primarily focused on deepfake detection, fall short due to the evolving sophistication of artificial intelligence methods and the emergence of what researchers term the "liar's dividend"—the ability for bad actors not only to spread fake content but also to dismiss authentic content as potentially fake. This dynamic, systematically analysed by Chesney and Citron (2019) in the California Law Review, highlights the erosion of epistemic trust when authenticity itself becomes contestable (Chesney and Citron, 2019; Pfeiffer and Krishna, 2025).

The limitations of current detection technologies are further compounded by the sheer volume and velocity of content being generated and shared online. Social media platforms, in particular, are fertile ground for the spread of deepfakes, as their algorithms are often designed to prioritise engagement and virality over accuracy and authenticity. The psychological phenomenon of FOMO, a pervasive anxiety that one might miss out on exciting or interesting events that others are experiencing, can further accelerate the spread of sensationalist content, as users are more likely to share information without verifying its authenticity in an effort to stay connected and engaged (Gong and Ren, 2023). The interplay between rapid deepfake evolution, limited detection capacity, and social amplification mechanisms such as FOMO creates fertile ground for misinformation, as illustrated by the Ozzy Osbourne case.

The field of deepfake detection has evolved rapidly, with researchers exploring diverse approaches ranging from traditional forensic techniques to advanced deep learning methods. A detailed review by Heidari et al. (2024) categorises detection methods into four primary domains: video detection, image detection, audio detection, and hybrid multimedia detection, each presenting unique challenges and opportunities. Verdoliva (2020) provides a foundational overview of media forensics in the deepfake era, noting that whilst traditional manipulation techniques such as splicing and copy-move forgery can be detected through pixel-level analysis, deepfakes generated by sophisticated neural networks present fundamentally different challenges. One promising avenue involves exploiting biological signals that are difficult for generative models to replicate. Ciftci, Demir, and Yin (2020) demonstrated that photoplethysmography (PPG) signals—subtle colour changes in facial skin caused by blood flow—can serve as an authenticity marker, as these physiological patterns are neither spatially nor temporally preserved in synthetic content. Their 'FakeCatcher' system achieved accuracies of 96% on Face Forensics and 91.07% on in-the-wild datasets by analysing biological signal maps through convolutional neural networks, suggesting that nature's own 'stamps' may provide more robust detection than artefact-based methods alone.

3. Case Study: The Deepfake Prophecy of Ozzy Osbourne's Death

The case of Ozzy Osbourne in July 2025 provides a compelling and tragic illustration of how deepfake technology can be weaponised to exploit public figures and manipulate public sentiment. In the weeks preceding his death, the 76-year-old rock legend, who had been publicly battling Parkinson's disease, became the subject of multiple viral deepfake videos that purported to show him speaking about his own impending demise. The timeline of events unfolded as follows: on 5 July 2025, Osbourne performed his "Back to the Beginning" farewell concert at Villa Park in Birmingham, England, reuniting with Black Sabbath for what would be his final show (Entertainment Weekly, 2025). Just four days later, on 9 July 2025, a TikTok clip surfaced featuring a synthetic voice-over claiming to be Osbourne discussing his final performance, stating, "This was my final performance. Uh, I have to say goodbye to this stage I've loved all my life. I'm truly sorry. I could only do this show strapped into a custom chair 'cause my body just couldn't handle it anymore. My Parkinson's has reached stage 5" (Lead Stories, 2025).[2] The video further claimed that Osbourne had donated all proceeds from the concert to Parkinson's research, a statement that was never corroborated by fact-based reporting.

The deepfake videos immediately fuelled speculation and conspiracy theories, with many social media users accepting the content as authentic. The clips were often presented in the context of a rumoured suicide pact between Osbourne and his wife, Sharon, a narrative that had its origins in historical interviews where Sharon Osbourne had discussed their shared belief in euthanasia under certain circumstances. In 2023, Sharon confirmed on the family's podcast that they still maintained an assisted suicide pact, stating they would go to Switzerland for euthanasia if either developed Alzheimer's disease (Yahoo Entertainment, 2025). The deepfake videos, therefore, did not emerge in a vacuum but rather tapped into and amplified existing public narratives and anxieties surrounding Osbourne's health. The situation prompted a swift and angry response from

[2] Archived here: https://web.archive.org/web/20250725165520/https://www.tiktok.com/@fametalks.01/video/7525036503290973454?lang=en

Osbourne's daughter, Kelly, who took to Instagram on 11 July 2025 to denounce the videos as malicious fabrications. In her public statement, she vehemently denied the rumours of her father's impending death and criticised the creators and distributors of the deepfakes, asking, "What is wrong with you?" (People Magazine, 2025). The incident was covered by multiple international news outlets, including People Magazine, Entertainment Weekly, and various social media platforms, indicating the widespread circulation of the deepfakes before Osbourne's actual death.

However, the deepfake phenomenon surrounding Osbourne did not end with Kelly's denial. The TikTok video posted on 9 July 2025 by the account @fametalks.01 continued to circulate, and was subsequently debunked by Lead Stories, a fact-checking organisation, on 26 July 2025. Lead Stories used VERA.ai detection software to analyse the video, and the analysis confirmed AI-generated segments with greater than 99% certainty and non-matching voice profiles, and the voice-over did not match any actual statements issued by Osbourne or corroborated by fact-based reporting (Lead Stories, 2025). The narration was identified as characteristic of AI text-to-speech programs, with no effort made to synchronise the audio with mouth movements in the video clips, further confirming its synthetic nature. The fact-check article noted that the video was a narrated PowerPoint-style slideshow with a series of clips and still images from the Villa Park show, and that the producer had no access to video of Osbourne saying these things.

Tragically, when Ozzy Osbourne passed away on 22 July 2025 (BBC News, 2025), just eleven days after Kelly's denial and thirteen days after the TikTok deepfake video was posted, the fabricated content took on a new and even more insidious dimension. Following Osbourne's death, the fabricated clips were retrospectively interpreted as prophetic, demonstrating deepfakes' capacity to manipulate grief and blur truth boundaries. The family was subsequently confronted with "disgusting" conspiracy theories, including false suicide claims that family friends publicly denounced as "vile" rumours (Watson.de, 2025). This phenomenon highlights a particularly dangerous aspect of deepfakes: their ability to be re-contextualised by real-world events, making them even more potent tools of misinformation. Social media discussions following Osbourne's death revealed widespread exposure to AI-generated tributes and memorial content. One such video, with 5.8 million views, exemplified the problem of "AI slop"—synthetic content designed solely for engagement, featuring an uncanny rendering of Osbourne with inappropriate music (Barlow, 2025). The Osbourne case demonstrates how deepfakes can not only create false narratives but also hijack and distort real ones, amplifying public grief and confusion in the process. It serves as a powerful reminder of the urgent need for more effective strategies to counter the spread of synthetic media, from improved detection and provenance tools to greater public awareness and critical media literacy.

Whilst this case study provides valuable insights into the real-world dynamics of deepfake misinformation, it is important to acknowledge certain limitations. As a single-case analysis focused on a celebrity figure with pre-existing health concerns and public narratives, the findings may not generalise to all deepfake scenarios, particularly those targeting private individuals or operating in different cultural contexts. Additionally, the retrospective nature of this analysis means that we cannot fully quantify the reach and impact of the deepfakes or measure the effectiveness of counter-narratives in real-time. Nevertheless, the Osbourne case offers a rich illustration of the temporal dynamics, psychological vulnerabilities, and detection challenges that characterise deepfake incidents more broadly.

4. The Human Element: Psychological Drivers of Misinformation

The viral spread of the Ozzy Osbourne deepfakes cannot be attributed solely to the sophistication of the technology itself. The incident also underscores the critical role of human psychology in the dissemination of misinformation. One of the key psychological drivers at play is FOMO, which is characterised by a persistent anxiety that one might miss out on rewarding experiences that others are having (Gong and Ren, 2023). In the context of social media, FOMO often manifests as a compulsive need to stay connected and up-to-date with the latest news and trends, leading to a higher propensity to share content without adequate verification. Research has shown that FOMO mediates the relationship between psychological distress and susceptibility to misinformation, with individuals experiencing higher levels of FOMO being more likely to believe fake news (Gong and Ren, 2023). FOMO interacts with the liar's dividend (discussed later in detail) (Pfeiffer and Krishna, 2025) to create a feedback loop between affective arousal and misinformation sharing. Social media platforms, with their constant stream of updates and notifications, provide an "efficient and low friction path" for information consumption, which can exacerbate FOMO and make users more susceptible to manipulation (Gong and Ren, 2023). The sensational and emotionally charged nature of the Osbourne deepfakes made them

particularly well-suited to exploit this psychological vulnerability, as users were driven to share the content to be part of the unfolding drama and to signal their awareness of the 'breaking news.'

Beyond FOMO, other cognitive and emotional factors contribute to the public's susceptibility to fake news. Research has shown that older individuals may be particularly vulnerable to misinformation, as they are often less likely to scrutinise suspicious content and may have lower levels of digital literacy (Gong and Ren, 2023). A cross-sectional study of earthquake survivors found that post-traumatic stress disorder (PTSD) was directly associated with believing fake news, and that this relationship was mediated by FOMO, with effects more pronounced in older people and present in females but not males (Gong and Ren, 2023). Whilst the Osbourne case does not involve a traumatic event in the traditional sense, the public's emotional investment in the rock star's health and the anticipation of his death created a heightened state of anxiety and vigilance that may have made individuals more susceptible to believing and sharing the deepfakes. The emotional content of misinformation also plays a significant role; people are more likely to believe and share false statements that appeal to their existing beliefs, biases, and emotions. In the Osbourne case, the deepfakes tapped into a pre-existing narrative of the rock star's declining health and the public's affection for him, creating a potent emotional cocktail that made the content more believable and shareable. The fact that the deepfakes were presented as 'prophecies' that were later 'fulfilled' by Osbourne's death further reinforced their emotional impact, making it even more difficult for some to dismiss them as fabrications.

The challenge of combating deepfake-driven misinformation is therefore not just a technical one; it is also a profoundly human one. It requires a deeper understanding of the psychological mechanisms that make us vulnerable to manipulation and the development of strategies to foster greater critical thinking and media literacy. This includes educating the public about the existence and capabilities of deepfake technology, teaching them how to identify the signs of a potential fake, and encouraging a culture of verification before sharing. As the line between reality and artifice becomes increasingly blurred, the ability to critically evaluate and question the information we encounter online will be an essential skill for navigating the digital world responsibly.

5. Ethical and Governance Challenges: The Consent Gap and the Liar's Dividend

The Ozzy Osbourne case study is a microcosm of the broader ethical and governance challenges posed by the rapid advancement of generative AI. At the heart of these challenges is the issue of consent. The creation of deepfakes, particularly those that impersonate real individuals, raises profound questions about the right to one's own likeness and voice. Current legal and ethical frameworks, which are largely built on a traditional understanding of consent, are proving to be inadequate in the age of AI (Pistilli and Trevelin, 2025). As researchers have pointed out, whilst an individual might consent to their data being used for a specific purpose, they cannot meaningfully consent to the myriad potential outputs that their data could be used to generate, a phenomenon described as the "consent gap" (Pistilli and Trevelin, 2025). This is particularly problematic when AI-generated content creates new and unforeseen forms of personal representation, often without the individual's knowledge or permission, and can be distributed globally in an instant. The question of whether AI can be truly consentful remains a pressing ethical concern, as the technology's capacity to generate synthetic representations far exceeds the scope of traditional consent mechanisms (Pistilli and Trevelin, 2025).

The sourcing of training data for AI models is another area fraught with ethical and legal ambiguity. Many generative models are trained on vast datasets scraped from the internet, which often include personal images, videos, and writings that are used without the explicit consent of the individuals concerned. This practice raises significant privacy concerns and has led to calls for greater transparency and accountability in the AI development pipeline. The development of robust governance frameworks is therefore essential to address these challenges. This includes the establishment of clear regulations regarding the use of personal data in AI training, as well as the implementation of policies that mandate the transparent labelling of synthetic media. In response to the growing threat of deepfakes, some jurisdictions are taking pioneering legislative action. Denmark, for instance, announced a proposal in June 2025 to amend its copyright laws, granting individuals rights over their biometric likenesses and positioning facial and vocal features as intellectual property to provide stronger legal recourse against AI-generated impersonations (The Guardian, 2025). This proposal has not yet been enacted. Similarly, Northern Ireland has taken steps towards criminalising the creation and sharing of deepfake pornography, reflecting a critical step towards addressing the severe psychological harm and gendered abuse associated with such content. Justice Minister Naomi Long stated in July 2025 that she wants deepfakes to become a criminal offence "sooner rather than later" (BBC News, 2025).

Beyond the issue of consent, the deepfake phenomenon has given rise to what researchers term the "liar's dividend"—the ability for bad actors not only to spread fake content but also to dismiss real footage as fake (Chesney and Citron, 2019; Pfeiffer and Krishna, 2025). This erosion of trust in authentic media represents a fundamental threat to democratic discourse and accountability. If seeing and hearing can no longer be believed, the very notion of truth is undermined, and public figures can evade accountability by falsely claiming that damaging but authentic content is fabricated (Chesney and Citron, 2019; Pfeiffer and Krishna, 2025). The Osbourne case illustrates this dynamic in reverse: whilst the deepfakes were fabrications, their retrospective 'confirmation' by Osbourne's actual death created a situation where some members of the public may have been more inclined to believe future synthetic content, further eroding the boundary between truth and falsehood. Policymakers must therefore prevent manipulated media from being used to undermine elections, disenfranchise voters, and distort public discourse (Brennan Center for Justice, 2023).

Initiatives like the Content Authenticity Initiative (CAI) and the Coalition for Content Provenance and Authenticity (C2PA) are pioneering technical solutions to address these challenges. The CAI, a cross-industry community including civil society, media, and technology companies, is promoting the adoption of Content Credentials, which function like a nutrition label for digital content (Content Authenticity Initiative, 2025). These credentials are based on open-source tools for verifiably recording the provenance of any digital media, including content made with generative AI, and focus on verifying the origin, integrity, and history of digital content. The C2PA specification reached version 2.1, reinforcing interoperability in provenance metadata (C2PA, 2025). Building upon C2PA's open standard, Pfeiffer and Krishna (2025) propose cryptographic hashing and metadata anchoring integrated into capture devices, optionally linked to eIDAS or SSI credentials, and immutably anchored on public blockchains (Ethereum Layer 2, Solana, Ardor) to form a provenance ledger capable of rapid verification and tamper resistance. Such blockchain-based media authentication offers multiple advantages, including strengthened content integrity, clear provenance, rapid verification capabilities, deterrence of misinformation, protection of intellectual property, and resilience across decentralised platforms.

However, as policy experts have noted, technical solutions alone are not sufficient. They must be complemented by a comprehensive regulatory approach that moves beyond a simplistic "synthetic or authentic" binary and addresses the full spectrum of risks associated with AI-generated media, from its potential to undermine democratic processes to its use in personal harassment and fraud (Tech Policy Press, 2025). Current approaches to regulating synthetic content may face a number of limitations and trade-offs, including concerns related to privacy and security, device integration complexities, scalability, verification infrastructure, and evolving legal and regulatory landscapes (Tech Policy Press, 2025; Pfeiffer and Krishna, 2025). To overcome these barriers, parallel initiatives in user and stakeholder education are crucial, emphasising media literacy, transparency, and global inclusivity (Pfeiffer and Krishna, 2025). Policymakers and the public must move beyond "synthetic or authentic" binaries when considering responses to AI-generated media, recognising that the challenge is not simply to distinguish between real and fake, but to build systems and cultures that can maintain trust and accountability in an age of ubiquitous synthetic media (Tech Policy Press, 2025).

Beyond technological solutions, the legal landscape surrounding deepfakes remains inadequate to address the multifaceted harms they inflict. Ramluckan (2024) argues that existing legal frameworks—including defamation, fraud, and intellectual property law—are ill-equipped to handle the unique characteristics of deepfake-enabled harm, particularly the reputational and psychological damage that occurs even when deepfakes are quickly debunked. The Osbourne case exemplifies this gap: whilst the deepfakes were identified and condemned, the emotional distress to the family and the erosion of public trust in authentic media persisted. Current legislation in most jurisdictions focuses narrowly on pornographic deepfakes or election interference, leaving vast categories of harmful content—such as fabricated health announcements or manipulated celebrity statements—largely unregulated. Ramluckan emphasises that effective legal responses must balance protection against harm with preservation of legitimate creative expression, a challenge that requires nuanced understanding of both the technology and its societal impacts.

6. Conclusion

The proliferation of deepfake technology represents a paradigm shift in the landscape of digital information, posing a formidable threat to public trust and the very notion of objective reality. The case of Ozzy Osbourne's 'death prophecy' serves as a sobering illustration of this threat, demonstrating how synthetic media can be used to manipulate public emotion, amplify misinformation, and exploit personal tragedy. The incident underscores the inadequacy of current detection technologies, which consistently fail to keep pace with the advancements in generative AI, leaving the public vulnerable to increasingly sophisticated and believable forgeries. The

psychological drivers of misinformation, such as FOMO, further exacerbate the problem, creating a fertile ground for sensationalist and unverified content to go viral. The case was widely discussed in policy forums and resembles incidents recorded in the OECD AI Incident Database (OECD AI Policy Observatory, 2025), highlighting the real-world consequences of deepfake technology and the urgent need for comprehensive responses.

The Osbourne case also reveals the temporal dynamics of deepfake misinformation. The first deepfake video circulated in early July 2025, with a TikTok clip appearing on 9 July. Kelly Osbourne's public denial followed on 11 July, but the deepfakes continued to circulate. When Osbourne died on 22 July, the deepfakes gained retrospective credibility, demonstrating how synthetic media can be re-contextualised by real events to become even more potent tools of misinformation. This timeline illustrates the challenge of combating deepfakes in real-time, as the speed of creation and dissemination often outpaces the capacity for detection and debunking. The proliferation of AI-generated tributes and memorial content following Osbourne's death further demonstrates how synthetic media can infiltrate and distort genuine expressions of public grief, creating an environment where authenticity itself becomes suspect.

Addressing the diverse challenges of deepfakes requires a concerted and multi-pronged approach that combines technical innovation, robust governance, and widespread public education. Technical solutions, such as the blockchain-based provenance and authentication standards being developed by the C2PA and proposed by researchers, offer a promising path towards a more transparent and accountable digital ecosystem (Content Authenticity Initiative, 2025; C2PA, 2025; Pfeiffer and Krishna, 2025). The integration of cryptographic hashing, secure metadata anchoring, and digital identity signatures into media-capturing devices, with authenticity markers immutably anchored to public blockchain infrastructures, could create an incorruptible provenance ledger that significantly mitigates the impact of deepfakes (Pfeiffer and Krishna, 2025). However, these technical solutions must be supported by clear and comprehensive policy frameworks that regulate the development and deployment of generative AI, with a particular focus on the ethical sourcing of training data, the mandatory labelling of synthetic media, and the protection of individuals' rights over their own biometric likenesses (Brennan Center for Justice, 2023; Pfeiffer and Krishna, 2025). Legislative initiatives such as Denmark's proposed amendments to copyright laws and Northern Ireland's consultation on criminalising deepfake pornography represent important steps in this direction (The Guardian, 2025; Department of Justice (NI), 2025).

Ultimately, however, the most potent defence against deepfake-driven misinformation lies in a well-informed and critically engaged public. Fostering digital and media literacy on a global scale is therefore an urgent imperative. By equipping individuals with the skills to critically evaluate the information they encounter online, to recognise the psychological vulnerabilities that make them susceptible to manipulation, and to understand the capabilities and limitations of AI technology, we can build a more resilient and discerning public sphere. The Osbourne case demonstrates how synthetic media can exploit personal tragedy to manipulate public sentiment. Addressing this requires the convergence of authentication technologies, legislative foresight, and a critically literate public equipped to navigate truth in a synthetic era

7. Implications for Cybersecurity Professionals

The Osbourne case yields several actionable insights for cybersecurity practitioners and information security teams.

- First, temporal awareness is critical: deepfakes may gain credibility retrospectively when real events appear to validate them, necessitating continuous monitoring even after initial debunking efforts. Security teams should implement event-triggered alert systems that flag resurgent synthetic content following related real-world developments.
- Second, the detection gap identified by Chandra et al. (2025)—where state-of-the-art models experience approximately 50% performance degradation on in-the-wild content—underscores the inadequacy of relying solely on automated detection tools. Organisations should adopt a defence-in-depth approach that combines technical detection with human forensic analysis, provenance verification, and rapid response protocols.
- Third, the psychological dimension of deepfake propagation, particularly FOMO-driven sharing behaviours, suggests that user education and awareness training should be integrated into organisational security programmes. Employees and stakeholders must be equipped to recognise manipulation tactics and resist the impulse to share unverified sensational content. Fourth, the case highlights the value of proactive authentication frameworks: organisations managing public-facing communications should consider implementing C2PA-compliant content credentials and blockchain-

based provenance tracking for official statements, particularly those concerning sensitive topics such as health, security, or financial matters. This creates a verifiable baseline against which deepfakes can be identified.

- Finally, cybersecurity professionals should engage with cross-sector collaboration initiatives such as the Content Authenticity Initiative and contribute to the development of industry standards for synthetic media labelling and detection. The arms race between generation and detection technologies requires collective intelligence sharing and coordinated response mechanisms. Incident response plans should explicitly address deepfake scenarios, including protocols for rapid verification, public communication, legal recourse, and platform coordination to remove malicious content. The convergence of technical sophistication, psychological manipulation, and regulatory gaps demands that cybersecurity practice evolve beyond traditional threat models to encompass the unique challenges posed by synthetic media in the post-truth era.

Ethics Declaration: We did not conduct any research involving user data. We did not receive any funding for writing this paper. Our motivation for writing this academic opinion paper about the Ozzy use case was intrinsic and maybe paranoid in regard to the future development of AI generated deepfakes

AI Declaration: We used the following tools: Consensus APP: for literature research, DeepL Write Pro: for spell-checking and ensuring consistency with British English throughout the text, CoPilot: Not for 'generating', but for challenging our text and thoughts. It was also used for formatting the references list in accordance with the conference guidelines.

This paper is dedicated to Ozzy Osbourne, the Prince of Darkness.

References

Amelinckx, A. (2025) 'The wild death rumors about Ozzy Osbourne that his daughter shut down', *Grunge*, 16 July. Available at: https://www.grunge.com/1914338/wild-death-rumors-ozzy-osbourne-daughter-kelly-osbourne-shut-down/ (Last accessed: 19 October 2025).

Barlow, G. (2025) 'This Ozzy Osbourne tribute video may have 5.8M views, but it just shows everything wrong with AI slop (and what the hell is that music?)', TechRadar, 24 July. Available at: https://www.techradar.com/ai-platforms-assistants/this-ozzy-osbourne-tribute-video-may-have-5-8m-views-but-it-just-shows-everything-wrong-with-ai-slop-and-what-the-hell-is-that-music (Last accessed: 19 October 2025).

BBC News (2025) 'Deepfakes to become criminal offence in NI "sooner rather than later"', *BBC News*, 22 July. Available at: https://www.bbc.com/news/articles/cwyxv9k28evo (Last accessed: 19 October 2025).

BBC News (2025) 'Ozzy Osbourne dies, weeks after farewell show', *BBC News*, 22 July. Available at: https://www.bbc.com/news/articles/cyvjd3p887qo (Last accessed: 19 October 2025).

Brennan Center for Justice (2023) 'Regulating AI deepfakes and synthetic media in the political arena', 5 December. Available at: https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena (Last accessed: 19 October 2025).

C2PA (2025) *C2PA Specification v2.1*. Available at: https://c2pa.org/specifications/specifications/2.1/specs/C2PA_Specification.html (Last accessed: 19 October 2025).

Chandra, N. A. et al. (2025) 'Deepfake-Eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024', *arXiv preprint arXiv:2503.02857*. Available at: https://arxiv.org/abs/2503.02857 (Last accessed: 19 October 2025).

Chesney, R. and Citron, D. K. (2019) 'Deep fakes: a looming challenge for privacy, democracy, and national security', *California Law Review*, 107(6), pp. 1753–1820. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954 (Last accessed: 19 October 2025).

Ciftci, U. A., Demir, I. and Yin, L. (2020) 'FakeCatcher: detection of synthetic portrait videos using biological signals', *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi: 10.1109/TPAMI.2020.3009287. Available at: https://arxiv.org/abs/1901.02212 (Last accessed: 19 October 2025).

Content Authenticity Initiative (2025) *About the Content Authenticity Initiative*. Available at: https://contentauthenticity.org (Last accessed: 19 October 2025).

Department of Justice (NI) (2025) 'Proposals to criminalise sexually explicit deepfake images'. Available at: https://www.justice-ni.gov.uk/consultations/proposals-criminalise-sexually-explicit-deepfake-images (Last accessed: 19 October 2025).

Entertainment Weekly (2025) 'Ozzy Osbourne says farewell, plays final concert with Black Sabbath to hometown crowd of 40,000, 6 July'. Available at: https://ew.com/ozzy-osbourne-reunites-with-black-sabbath-for-final-concert-11767172 (Last accessed: 19 October 2025).

Gong, C. and Ren, Y. (2023) 'PTSD, FOMO and fake news beliefs: a cross-sectional study of Wenchuan earthquake survivors', *BMC Public Health*, 23(1), p. 2213. doi: 10.1186/s12889-023-17151-z.

Heidari, A., Jafari Navimipour, N., Dag, H. and Unal, M. (2024) 'Deepfake detection using deep learning methods: a systematic and comprehensive review', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2), e1520. doi: 10.1002/widm.1520.

Lead Stories (2025) 'Fact Check: The viral Ozzy "My final performance… every dime went to Parkinson's research" video is NOT authentic', 26 July. Available at: https://www.yahoo.com/news/articles/fact-check-viral-ozzy-final-032503404.html (Last accessed: 19 October 2025).

Lundberg, E. and Mozelius, P. (2025) 'The potential effects of deepfakes on news media and entertainment', *AI & SOCIETY*, 40, pp. 2159–2170. doi: 10.1007/s00146-024-02072-1.

OECD AI Policy Observatory (2025) 'AI-Generated Deepfake Video Falsely Claims Ozzy Osbourne Is Dying', *OECD AI Incidents Monitor*, 12 July. Available at: https://oecd.ai/en/incidents/2025-07-12-45a3 (Last accessed: 19 October 2025).

PCMag (2025) 'I tested out Google's Veo 3 AI video generator. The internet is not prepared', 10 June. Available at: https://www.pcmag.com/opinions/i-tested-out-googles-veo-3-ai-video-generator-the-internet-is-not-prepared (Last accessed: 19 October 2025).

People (2025) 'Kelly Osbourne debunks rumors about dad Ozzy Osbourne's health and clarifies "He's not dying"', 11 July. Available at: https://people.com/kelly-osbourne-debunks-rumors-about-ozzy-osbourne-health-clarifies-hes-not-dying-11770973 (Last accessed: 19 October 2025).

Pfeiffer, A. and Krishna, N. (2025) 'Restoring trust in the age of deepfakes: a blockchain-based proposal – the global challenge of AI-manipulated media', *MAD Opinions*, University for Continuing Education Krems, 30 July. doi: 10.48341/dx7k-e072.

Pistilli, G. and Trevelin, B. (2025) 'Can AI be consentful?', *arXiv preprint arXiv:2507.01051*. Available at: https://arxiv.org/abs/2507.01051 (Last accessed: 19 October 2025).

Ramluckan, T. (2024) 'Deepfakes: the legal implications', in *Proceedings of the 19th International Conference on Cyber Warfare and Security, ICCWS 2024*, pp. 282–288. Available at: https://papers.academic-conferences.org/index.php/iccws/article/view/2099 (Last accessed: 19 October 2025)

Tech Policy Press (2025) 'Synthetic media policy: provenance and authentication', 1 May. Available at: https://techpolicy.press/synthetic-media-policy-provenance-and-authentication-expert-insights-and-questions (Last accessed: 19 October 2025).

The Guardian (2025) 'Denmark to tackle deepfakes by giving people copyright over their faces', 27 June. Available at: https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence (Last accessed: 19 October 2025).

Verdoliva, L. (2020) 'Media forensics and deepfakes: an overview', *arXiv preprint arXiv:2001.06564*. Available at: https://arxiv.org/abs/2001.06564 (Last accessed: 19 October 2025).

Watson.de (2025) 'Ozzy Osbourne: Fake-Theorien zu seinem Tod – Familie ist "angewidert"', 29 July. Available at: https://www.watson.de/unterhaltung/stars/788233724-ozzy-osbourne-fake-theorien-zu-seinem-tod-familie-ist-angewidert (Last accessed: 19 October 2025).

Yahoo Entertainment (2025) 'Sharon and Ozzy Osbourne had decades-long assisted suicide pact', 22 July. Available at: https://www.yahoo.com/entertainment/articles/sharon-ozzy-osbourne-had-decades-222826539.html (Last accessed: 19 October 2025).

How Did They get Aboard my Ship? Analysing Vessel Cyber Incidents Using the Cyber Kill Chain

Jeroen Pijpker and Stephen McCombie
NHL Stenden University of Applied Sciences, Leeuwarden, The Netherlands

jeroen.pijpker@nhlstenden.com
stephen.mccombie@nhlstenden.com

Abstract: The Global Maritime Transportation System (GMTS) is increasingly reliant on digitized and interconnected systems. This trend also applies to one of the most essential elements of the GMTS, vessels themselves. The Maritime Cyber Attack Database (MCAD) contains listings of maritime cyber incidents from public sources, including those involving vessels. Utilizing MCAD in this paper, we categorize and analyse these specific listings related to vessels. Vessels have complex networks of Information Technology (IT) and Operational Technology (OT). In this paper, we further categorize and systematically analyse selected cyber-attacks against vessels with the Lockheed Martin Cyber Kill Chain. This work also shows a significant rise in attacks against vessels since 2016. These case studies help industry more clearly understand the specific tactics and techniques used in cyber-attacks on vessels.

Keywords: Maritime cybersecurity, Cyber kill chain, MCAD, GMTS

1. Introduction

The Maritime Sector is increasingly dependent on digital systems, this is also the case for perhaps the key element, vessels themselves. Vessels are becoming more reliant on digitized and interconnected systems to support the different onboard systems. Vessels are also increasingly using systems that rely on an Internet connection to operate. Thus, extending the attack surface of vessels and exposing new vulnerabilities in these cloud-based connections.

Modern vessels integrate IT and OT systems into a single ecosystem. The OT systems on a ship control essential functions, such as engine management, radar, and dynamic positioning. Many of these systems were designed with safety and reliability in mind, but not cybersecurity. The increasing digitalization of maritime operations has introduced vulnerabilities that adversaries can exploit in the vessel's complex network of IT and OT.

This work illustrates how cyber-attacks against vessels are crafted according to the steps of the Cyber Kill Chain. The Lockheed Martin Cyber Kill Chain framework is applied to model the sequence of events for selected attacks from the Maritime Cyber Attack Database (MCAD) relating to vessels (Maritime IT Security Research Group, 2025).

1.1 Related Work

The research area of maritime cybersecurity is relatively new and niche. That said, there is some focus in the literature about cyber-attacks on the maritime sector and vessels in particular.

The paper by Simola et al. 2023 (Simola, Pöyhönen and Lehto, 2023) discusses different cyber-attacks against the GMTS utilizing the MITRE ATT&CK framework to map threat actors' objectives. The paper provides an overview of 13 attacks on ports from 2013 to 2023 in most cases targeted by ransomware. Iqbal and Khan (Iqbal and Khan, 2021) discuss different cyber-attacks against GMTS and provide an overview of 13 attacks. Of these 13 attacks, two of them involved vessels. In one case, a cyber-attack was impacting the shipboard network, and the case was investigated by the US Coast Guard.

Other research focuses on the specific methods and possible defense measures. The paper written by Mraković and Vojinović (Mraković and Vojinović, 2019) discusses cybersecurity threats and attacks to the GMTS, highlighting 9 incidents including "NotPetya" malware incident which impacted maritime companies COSCO, Maersk, and others.

Farah et al. (Ben Farah *et al.*, 2022) gives a summary of 16 cyberattacks on the GMTS. One case of a spoofing attack carried out by a research team was just a proof of concept rather than a real incident. Five of the attacks involved vessels. Those vessel incidents included navigation systems attacks, a ransomware attack and a phishing attack. The paper written by Schwarz et al. (Schwarz, Marx and Federrath, 2021) provides a comprehensive analysis of information security incidents in the GMTS over the previous 20 years. It covers 90 publicly reported attacks and 15 proof of concepts, categorizing them into different threat categories. Many of the attacks involved vessels.

Meland et al. (Meland *et al.*, 2021) identified and analysed 46 cyber incidents in the GMTS, making a top-ten list of maritime cyber threats, ranging from threats in IT and OT systems, to economic fraud and manipulation of GNSS signals. Number one in the top-ten list is the exposed shipping company/carrier IT-systems. Oruc (Oruc, 2020) emphasizes the impact of cyber-attacks on the maritime industry due to its widespread use of information technology (IT) and operational technology (OT). Twenty-two maritime cybersecurity incidents were analysed that occurred between 2011 and 2019. Five of these were attacks on vessels. The vessels were targeted with malware, GPS jamming/spoofing and other undefined navigation attacks. Cichocki (Cichocki, 2023) discusses state-sponsored and organized crime threats to maritime transportation systems of the Ukraine. The focus is more on cyber warfare not directly related to maritime threats. In a study conducted by Jones et al. (Tam, Jones and Papadaki, 2016) they presented several scenarios to demonstrate possible cyber-attacks on vessels.

2. Background

2.1 Why is the GMTS so Critical and Vulnerable

The role of GMTS in the global economy is significant with over 80 percent of the world's cargo transported by ship (Bronk and deWitte, 2020) and representing 70 percent of global trade by value (Loomis *et al.*, 2021). At the same fleets are aging and their technology is aging with them and thus more vulnerable to cyber-attacks. 38 percent of oil tankers and 59 percent of general cargo ships are older than 20 years (Tam and Jones, 2018). Supply chains are increasingly impacted by cyber-attacks. This is particularly stark in recent years with a four fold increase in supply chain cyber-attacks in 2021 compared to 2020 (Kessler and Shepard, 2020). The GMTS is a key part of global supply chains.

Many vessels have vulnerable legacy systems and using Operating Systems (OS) that themselves are no longer under support. These legacy systems have many published vulnerabilities with exploits available. Even newer systems are delivered on outdated OS due to the costs of re-engineering to suit more modern OS versions. Physical access to vessels is limited for remediation. As a result, vessels are often only available for updating and patching due to small windows they spend in their home port. There is also limited IT expertise on board and functions related to IT often are only part-time responsibility for those involved. Naturally, cybersecurity expertise is usually not present on board. VSAT systems used for vessels are also poorly maintained and have been targeted by various threat actors given their critical role in vessels. Network segmentation is primitive if it exists at all. While there is some monitoring of security events taking place it is the exception far from the rule.

2.2 Methodology

The methodology for this work was to firstly to identify and collect vessel cyber incidents where enough detail exists to categorize and summarize each incident according to the Cyber Kill Chain. MCAD is the primary data source utilized for this analysis.

2.2.1 MCAD

The authors and other researchers at NHL Stenden University of Applied Sciences in the Netherlands have compiled MCAD since 2021. MCAD catalogues cyber-attacks targeting the GMTS. Currently, MCAD has catalogued over 440 cyber incidents against GMTS dating back to 2001. The MCAD database is online and publicly available. MCAD contains listings of actual maritime cyber incidents from public sources, including those involving vessels. MCAD defines a cyber incident as a discrete malicious attack with a cyber element, perpetrated by a particular threat actor against one or more victims and causing significant impact on one or more victims, possibly over an extended period. A cyber incident is eligible for inclusion in MCAD if it involves at least one component of the GMTS, which is defined as a ''system of systems'' that encompasses several key components essential to global trade and mobility. At its core are port operations, where the loading, unloading, and transshipment of goods take place, and shipping lines, which facilitate the movement of cargo and passengers on designated routes. This system also integrates vessel traffic control systems, which monitor and direct maritime traffic to ensure safe and efficient operations, and vessels themselves—ranging from container ships to tankers and passenger liners—that transport goods and people across oceans. MCAD has 67 cyber incidents relating to vessels. The conversion to cyber kill chain included all the relevant information from the various sources and listed in MCAD.

3. Cyber Attack on Vessels

Per Figure 1, we have extracted all maritime cyber incidents relating to vessels from MCAD and there are 67 cyber incidents falling into 7 category types. The top three are related to navigation AIS/GPS Spoofing (25), Going

Dark (15) and GPS Jamming (12). Then two categories related to malware, Ransomware (4) and Other Malware (3). Lastly Hacking (6) and Other (2). The number of cyber-attacks has been significantly increasing over time. Between 2001-2015 (14 years) there were only 7 incidents while between 2016-2025 (9 years) there were 60.

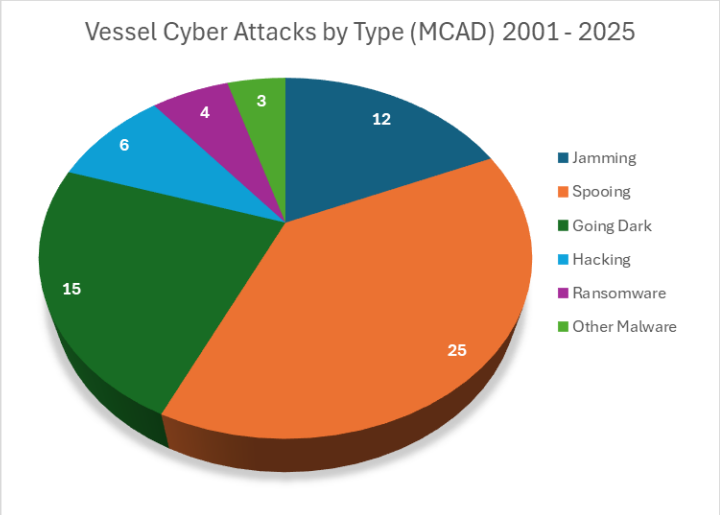


Figure 1 The figure represents MCAD cyber incidents from 2001-2025 relating to vessels.

4. Cyber Kill Chain

Cyber Kill Chain was developed by Lockheed Martin in 2011 (Hutchins, Cloppert and Amin, 2011). The Cyber Kill Chain is part of the Intelligence Driven Defense model for identifying and preventing cyber intrusion activity, identifying what a threat actor must complete in order to achieve their goals. The cyber kill chain helps in analyzing how an attacker must successfully and consecutively progress through all seven stages to successfully complete a cyber-attack. The seven stages of the cyber kill chain are defined as Reconnaissance, Weaponization, Delivery, Exploitation, Installation, Command and Control (C2) and Actions on Objectives. In a study conducted by Bahrami et al. (Bahrami *et al.*, 2019), a Cyber Kill Chain taxonomy is presented based on a large-scale analysis of attacks from 40 APT campaigns. The taxonomy captures the TTPS utilized by attackers. This taxonomy can help by choosing the right terminology in the kill chain analysis.

5. Case Studies

5.1 Selection of Case Studies

The cyber incidents are described based on the information in the MCAD (Maritime IT Security Research Group, 2025). The selected incidents are shown in Table 1. For each selected incident, there is a table mapping the event to the Cyber Kill Chain Table 2 - Table7. The selection of these six specific case studies was made based on the availability of information required for the Cyber Kill Chain framework. The selected cases (Table 1) range from physical breaches through USB and remote phishing to signal spoofing.

Table 1: Selected vessel cyber incidents from MCAD for analysing

Year	Incident	Impact area	Incident location	Method	References
2018	New-build dry bulk ship in port hit by ransomware attack	Vessel	not available	MalwareRansomware	(Safety4Sea, 2018; SecureWorld, 2018; ZDNet, 2018; Akpan, 2022)
2019	Two ships hit by ransomware attack	Vessel	not available	Malware Ransomware	(Abrams, 2018; Meland *et al.*, 2021; Akpan *et al.*, 2022)
2019	Chinese Hackers Targeting Ships Across Europe with Malware on USB Sticks	Vessel	Norway, Greece, Netherlands	Malware	(News, 2024; SupplyChainBrain, 2024; The Maritime Executive, 2024)
2025	Hacktivists Disrupt Communications of Iranian Oil Tankers	Vessels	Persian Gulf	VSAT hacking	(Cydome, 2025)

Year	Incident	Impact area	Incident location	Method	References
2012	USS Harry S Truman Insider Attack	Nuclear Aircraft Carrier	Atlantic Ocean	SQL Injection	(Naked Security, 2014; U.S. Attorney's Office, Northern District of Oklahoma, 2014; Zetter, 2014)
2018	Stena Impero GPS Spoofing	Vessel	Strait of Hormuz	GPS Spoofing	(Androjna and Perkovič, 2021)

5.2 Case Study 1: New-build dry Bulk Ship in Port hit by Ransomware Attack

In 2018, a bunker surveyor on a new build ship inserted a USB thumb drive into an onboard computer and unwittingly introduced malware onto the ship's administrative network which executed a ransomware attack. (Safety4Sea, 2018, 2018; SecureWorld, 2018; ZDNet, 2018; Akpan, 2022). The malware went undetected until a cyber assessment was conducted on the ship later after the crew reported a computer issue affecting the business network. The ship owner paid the ransom. The mapping of the Cyber Kill Chain to Case Study 1 is displayed in Table 2.

Table 2: Case Study 1 Mapped to Cyber Kill Chain

Stage	Mapping to the incident
Reconnaissance	The attacker carried out none due to limited surveillance. This can be seen as opportunistic infection via third-party USB device.
Weaponization	The malware was preloaded on the USB drive, likely due to an earlier infection on the surveyor's computer.
Delivery	Delivery of the malware took place through the use of the USB device into the computer in the engine room.
Exploitation	The malware had the opportunity to spread because the system allowed the use of untrusted removable media devices, and there is a likelihood of auto-executing upon insertion due to outdated OS or autorun settings.
Installation	The malware was able to install itself on the administrative network, spreading to the connected business network.
Command and Control (C2)	The cyber incident does not mention anything related to communication. But there is always the possibility that the malware has connected to a command and control, for example, for further instructions.
Actions on Objectives	The business network was disrupted, leading to delays and the ransom was paid. The recovery costs were totalled at around $100,000.

5.3 Case Study 2: Two ships hit by Ransomware Attack

In 2019, two ships were hit by simultaneous ransomware attacks. The infection came as a macro-enabled Word document attached to an email, and multiple workstations on the administrative networks were affected (Abrams, 2018; Meland *et al.*, 2021; Akpan *et al.*, 2022). When these emails were opened, they installed the AZORult information-stealing Trojan and the Hermes 2.1 ransomware onto the recipient's computer. All versions of Hermes (2.0 \& 2.1) Ransomware avoid infecting Russian, Ukrainian and Belarusian netizens. The result of mapping the Cyber Kill Chain to Case Study 2 is displayed in Table 3.

Table 3: Case Study 2 Mapped to Cyber Kill Chain

Stage	Mapping to the incident
Reconnaissance	Probably minimal reconnaissance from the attacker in this situation. The attackers probably performed broad targeting using an email address list of (maritime) companies or staff
Weaponization	Word document (named: Invoice.doc) was weaponized with the AZORult Trojan and the Hermes 2.1 ransomware
Delivery	Delivered via phishing spam campaign with email subject "Invoice Due". The email contained a password protected malicious attachment.
Exploitation	The exploitation took place through user interaction. 1. Entering the password ('1234') 2. Clicking "Enable Content" allowed macros to be executed.
Installation	The AZORult Trojan (azo.exe) downloaded and started executing. Next step was the downloading of the Hermes 2.1 ransomware (hrms.exe).

Stage	Mapping to the incident
Command and Control (C2)	AZORult established an outbound communication channel for callback traffic. Possible exfiltration of credentials, session info, or host data.
Actions on Objectives	The Hermes 2.1 encrypted data on the target system. It showed a ransom note (DECRYPT_INFORMATION.html). The Hermes 2.1 ransomware avoided infecting Russian, Ukrainian, and Belarusian netizens.

5.4 Case Study 3: Chinese Hackers Targeting Ships Across Europe with Malware on USB Sticks

In early 2024, Mustang Panda, a Chinese threat actor, targeted cargo shipping companies in Europe using Korplug loaders (ESET Research, 2024; News, 2024; SupplyChainBrain, 2024; The Maritime Executive, 2024). These loaders, dropped from USB drives with suspicious filenames, compromised systems in Norway, Greece, and the Netherlands. Some malware samples were blocked, but others had altered signatures and utilized DLL hijacking. The result of mapping the Cyber Kill Chain to Case Study 3 is displayed in Table 4.

Table 4: Case Study 3 Mapped to Cyber Kill Chain

Stage	Mapping to the incident
Reconnaissance	Targeted European cargo shipping companies. Specific victims in Norway, Greece, and the Netherlands.
Weaponization	Korplug (PlugX) malware was packaged with custom loaders and suspicious filenames. DLL hijacking was used to bypass defences.
Delivery	The malware was delivered through an infected USB thumb drive.
Exploitation	The execution of the payload started when files on the USB were opened. The attack relied on user interaction or autorun setting. Through DLL hijacking exploited further system weaknesses.
Installation	Persistence was achieved through backdoors (e.g. PlugX), this enabled remote access to the infected machines.
Command and Control (C2)	Some of the acquired malware samples maintained communication with external servers controlled by the attacker. This enabled remote command and data exfiltration.
Actions on Objectives	Espionage, surveillance, with the exfiltration of sensitive data and operational data.

5.5 Case Study 4: Hacktivists Disrupt Communications of Iranian Oil Tankers

In March 2025 cybersecurity company Cydome reported that the anti-Iranian government hacktivist group "Lab Dookhtegan" ("sealed lips" in Farsi) claimed to have disrupted communications on over 100 oil tankers associated with Iranian government-linked company (Cydome, 2025). While Lab Dookhtegan did not disclose the specific methods used, it is believed they exploited vulnerabilities in the ships' satellite communication systems, such as VSAT terminals. These systems are known to be susceptible to cyber-attacks, especially when default passwords remain unchanged. From those systems, they can take complete control over all communications of the vessel and even spread out to the IT and OT systems. The result of mapping the Cyber Kill Chain to Case Study 4 is displayed in Table 5.

Table 5: Case Study 4 Mapped to Cyber Kill Chain

Stage	Mapping to the incident
Reconnaissance	Lab Dookhtegan identified Iranian government linked tankers probably through AIS data, IP ranges of SATCOM terminals (Shodan), or vendor documentation.
Weaponization	The attackers exploited the insecure SATCOM systems (e.g. default credentials, exposed web interfaces).
Delivery	Gained access through remote accessible SATCOM devices. Probably using known IP addresses and (default) login credentials.
Exploitation	Exploitation of default/weak passwords, remote shell access to SATCOM.
Installation	Possible creation of backdoors to maintain access. For example, installation of scripts, altering of startup job, or SATCOM config files.
Command and Control (C2)	Remote control through custom C2 channel.
Actions on Objectives	Disrupted vessel communication on 100+ ships. Ships are isolated at sea. Possibility of lateral movement to onboard IT/OT systems.

5.6 Case Study 5: USS Harry S Truman Insider Attack

In June 2012, a land based IT system of the United States Navy was impacted by a SQL hacking attack (Naked Security, 2014; Zetter, 2014; United States Department of Justice, 2015). The impact of hacking on the Navy-SWM system was theft of confidential information of approximately 220,000 Navy service people who were being transferred. As a result, Navy-SWM system was shut down and never resumed operation. The incident caused a loss to the Navy of approximately USD$514,000. The US Naval Criminal Investigation Service (NCIS) worked out the attack was executed from onboard a US Navy nuclear aircraft carrier, the USS Harry S. Truman, which was at sea at the time. The threat actor they determined was a serving US Navy sailor; Nicholas Paul Knight whose job onboard was to run the IT Network for the nuclear reactor. The surprising thing is not the attack on the US Navy system but the fact it was executed from aboard a US Navy ship. Thus, the compromise most significantly was of the USS Harry S. Truman by an insider. The result of mapping the Cyber Kill Chain to Case Study 5 is displayed in Table 6.

Table 6: Case Study 5 Mapped to the Cyber Kill Chain

Stage	Mapping to the incident
Reconnaissance	Knight was a Navy systems administrator aboard the USS Harry S. Truman. As an administrator, Knight had privileged internal access and knowledge about the onboard and Navy systems.
Weaponization	SQL injection was used to hack the US Navy's SWM database.
Delivery	SQL code injection took place within the Navy-SWM system.
Exploitation	SQL Injection enabled unauthorized queries which were able to extract personnel records form Navy-SWM database.
Installation	No known installation of malware.
Command and Control (C2)	Threat actor had admin access to his own systems and compromised access to the victim system. So, command and control was direct not via a botnet or other intermediary system.
Actions on Objectives	Personal data from over 220,000 service members was stolen and posted online. The Navy-SWM platform was prematurely shutdown, over 700 deployed overseas Service Members could not access logistical support for transfers for more than 10 weeks. Team Digi7al's Navy-SWM hack caused a loss to the US Navy of approximately USD$514,000.

5.7 Case Study 6: Stena Impero GPS Spoofing

In July 2019, the British-flagged tanker Stena Impero fell victim to a GPS spoofing incident in the Strait of Hormuz, Iran, resulting in the vessel violating Iranian territorial waters (Androjna and Perkovič, 2021). Subsequently, the vessel was seized by Iran's Revolutionary Guards and it has been reported that Iran and/or Russia were involved in purposefully spoofing the ship, so it strayed into Iranian waters. The result of mapping the Cyber Kill Chain to Case Study 6 is displayed in Table 7.

Table 7: Case Study 6 Mapped to the Cyber Kill Chain

Stage	Mapping to the incident
Reconnaissance	Target vessel, identifying its routes and possible geopolitical opportunities.
Weaponization	Spoofed GPS signal and possible AIS data spoofing/tampering
Delivery	Broadcasting of fake navigation signals
Exploitation	The shipboard equipment receives fake signals.
Installation	Deception becomes persistent. Spoofed location
Command and Control (C2)	Continues signal GPS spoofing
Actions on Objectives	Vessels enters Iranian waters and is taken over.

6. Discussion

The six case studies illustrate the varying nature of cyber-attacks on vessels. Applying the Cyber Kill Chain to the selected case studies indicates that vessels can be compromised via conventional IT attack vectors rather than via zero-day exploits. Reconnaissance is minimal for most attacks, except in geopolitically motivated incidents with the Stena Impero. Most attacks are opportunistic; attackers instead rely on broad targeting or accidental infection. Analysis of the Reconnaissance and weaponization phases shows that, for the selected case studies, threat actors do not need to develop specialized, maritime-specific code to breach a vessel. They can weaponize

standard available malware, such as the AZORult Trojan or generic ransomware, to carry out their attack. Looking at the delivery methods used in the attacks, we see both opportunistic infections and targeted state-sponsored campaigns that utilize USB drives to bridge air-gapped systems. Phishing remains an effective method. Phishing is also in the GMTS an effective method, primarily because it exploits the human element in the exploitation phase. The success of the techniques used in the exploitation phase depends on the still existing legacy systems on board some older vessels. Outdated features, such as autorun for USB devices, still exist. Examining the installation phase showed in the different case studies that attackers rarely require sophisticated persistence mechanisms. In some case studies, the malware successfully executed because onboard systems permit untrusted code to run, lack application whitelisting, and operate on legacy configurations. Command and Control (C2) phase showed two different patterns. Malware was able to establish an outbound connection due to the limited logging and monitoring on the vessels. The other, more advanced incidents include VSAT compromises, which enabled remote access through misconfigured satellite communications systems. So the lack of visibility means that C2 is mostly undetected. When examining the Actions on Objectives, incidents range from financial crime to geopolitical disruption. In the case of the Stena Impero, GPS spoofing caused the vessel to enter Iranian waters. Lack of network segmentation enabled malware to move from the engine room to the administrative network. The case studies examined highlight recurring technical weaknesses: reliance on removable media, minimal network segmentation, legacy systems, and insufficient access control. Summarizing the weak points, the GMTS must, for example, enforce removable media policies, network segmentation and harden systems against unauthorized access.

7. Conclusion

This work explored the mapping of maritime cybersecurity incidents from MCAD to the cyber kill chain. These maritime cybersecurity incidents are publicly documented and have multiple sources. The analysis of vessel specific cyber incidents with the Cyber Kill Chain reveals several behavioural patterns. The case studies used in this work range from opportunistic infections to highly sophisticated, targeted campaigns. This work illustrates that attackers are able to exploit a combination of technical weaknesses and human factors to gain access to shipboard systems. The Cyber Kill Chain proves to be a valuable tool for analyzing maritime cybersecurity incidents. The discussion also highlighted the persistence of legacy systems onboard and unpatched vulnerabilities, allowing conventional IT threats to get to the OT segment of the network. A key conclusion from the research is the fact that it is difficult within the GMTS to be able to detect and respond to threats in real time. This comes because of the nature of the vessels. Poor network segmentation, outdated systems, and a lack of cyber hygiene aboard a vessel can play a crucial role in cyber incidents. Further work will explore the use of a ship honeynet (McCombie and Pijpker, 2022; Pijpker and McCombie, 2023) to better understand the tactics, techniques and procedures (TTPs) used in cyber-attacks on vessels. This can help better understand the attacker's behaviours and can in the future also be mapped to frameworks like MITRE ATT&CK. The efforts will contribute to a more resilient maritime cybersecurity posture.

AI Declaration: Artificial intelligence tools (e.g., Grammarly) were used to assist in editing and improving the language of this paper. The authors are fully responsible for the content, analysis, and conclusions.

Ethics Declaration: This research did not require ethical clearance as it did not involve human participants, personal data, or animal experiments.

References

Abrams, L. (2018) "Beware of Spam with Fake Invoices Pushing Hermes 2.1 Ransomware and AZORult," *BleepingComputer* [Preprint]. Available at: https://www.bleepingcomputer.com/news/security/beware-of-spam-with-fake-invoices-pushing-hermes-21-ransomware-and-azorult/.

Akpan, E. (2022) "Cyberattacks on Maritime Transportation Systems: A Threat to Global Supply Chains," *Journal of Marine Science and Engineering*, 2(1). Available at: https://www.mdpi.com/2673-8732/2/1/9/pdf.

Akpan, F. *et al.* (2022) "Cybersecurity Challenges in the Maritime Sector," *Network*, 2(1), pp. 123–138. Available at: https://doi.org/10.3390/network2010009.

Androjna, A. and Perkovič, M. (2021) "Impact of Spoofing of Navigation Systems on Maritime Situational Awareness," *Transactions on Maritime Science*, 10(2), pp. 361–373. Available at: https://doi.org/10.7225/toms.v10.n02.w08.

Bahrami, P.N. *et al.* (2019) "Cyber kill chain-based taxonomy of advanced persistent threat actors: Analogy of tactics, techniques, and procedures," *Journal of information processing systems*, 15(4), pp. 865–889.

Ben Farah, M.A. *et al.* (2022) "Cyber Security in the Maritime Industry: A Systematic Survey of Recent Advances and Future Trends," *Information*, 13(1), p. 22. Available at: https://doi.org/10.3390/info13010022.

Bronk, R. and deWitte, P. (2020) "Maritime Cybersecurity: Meeting Threats to Globalization's Great Conveyor," in. Available at: https://doi.org/10.24251/HICSS.2020.240.

Cichocki, R. (2023) "State-Sponsored and Organized Crime Threats to Maritime Transportation Systems in the Context of the Attack on Ukraine," *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, 17(3), pp. 717–721. Available at: https://doi.org/10.12716/1001.17.03.24.

Cydome (2025) *Lab Dookhtegan cyber attack on Iranian oil tankers disrupts operations*. Available at: https://cydome.io/lab-dookhtegan-cyber-attack-on-iranian-oil-tankers-disrupts-operations/.

ESET Research (2024) *APT Activity Report Q4 2023–Q1 2024: Iran-aligned Cyberattacks – Rise in Disruptive Operations*. ESET. Available at: https://web-assets.esetstatic.com/wls/en/papers/threat-reports/eset-apt-activity-report-q4-2023-q1-2024.pdf.

Hutchins, E.M., Cloppert, M.J. and Amin, R.M. (2011) "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains," *Leading Issues in Information Warfare & Security Research*, 1(1), p. 80.

Iqbal, Z. and Khan, M.K. (2021) "Maritime cybersecurity: Vulnerabilities and counter measures," *Journal of Contemporary Studies*, 9(II), pp. 42–58.

Kessler, G.C. and Shepard, S.D. (2020) *Maritime cybersecurity: a guide for leaders and managers*. Gary C. Kessler and Steven D. Shepard.

Loomis, W. *et al.* (2021) "Raising the Colors: Signaling for Cooperation on Maritime Cybersecurity."

Maritime IT Security Research Group (2025) "Maritime Cyber Attack Database (MCAD)," *NHL Stenden University of Applied Sciences* [Preprint]. Available at: https://maritimecybersecurity.nl/.

McCombie, S.J. and Pijpker, J. (2022) "A Ship Honeynet Project to Collect Data on Cyber Threats to the Maritime Sector."

Meland, P.Há. *et al.* (2021) "A Retrospective Analysis of Maritime Cyber Security Incidents," *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, 15(3), pp. 519–530. Available at: https://doi.org/10.12716/1001.15.03.04.

Mraković, I. and Vojinović, R. (2019) "Maritime Cyber Security Analysis – How to Reduce Threats?," *Transactions on Maritime Science*, 8(1), pp. 132–139. Available at: https://doi.org/10.7225/toms.v08.n01.013.

Naked Security (2014) *Team Digi7al: US Navy hacker sentenced to 2 years in jail*. Available at: https://web.archive.org/web/20211018083537/https://nakedsecurity.sophos .com/2014/10/29/teamdigi7al-us-navy-hacker-sentenced-to-2-years-in-jail/.

News, N.B.C. (2024) "China-linked group uses malware to try to spy on commercial shipping, new report says," *NBC News* [Preprint]. Available at: https://www.nbcnews.com/news/world/china-linked-group-malware-spy-commercial-shipping-cargo-report-eset-rcna152129.

Oruc, A. (2020) "Claims of State-Sponsored Cyberattack in the Maritime Industry."

Pijpker, J. and McCombie, S.J. (2023) "A Ship Honeynet to Gather Cyber Threat Intelligence for the Maritime Sector," in *2023 IEEE 48th Conference on Local Computer Networks (LCN)*. *2023 IEEE 48th Conference on Local Computer Networks (LCN)*, pp. 1–6. Available at: https://doi.org/10.1109/LCN58197.2023.10223347.

Safety4Sea (2018) "BIMCO Guidelines on Cyber Security Onboard Ships." Available at: https://safety4sea.com/wp-content/uploads/2018/12/BIMCO-Guidelines-on-cyber-security-onboard-ships-2018_12.pdf.

Schwarz, M., Marx, M. and Federrath, H. (2021) "A Structured Analysis of Information Security Incidents in the Maritime Sector." arXiv. Available at: http://arxiv.org/abs/2112.06545 (Accessed: January 31, 2024).

SecureWorld (2018) "Ships at Sea: More Ways to Hack Them." Available at: https://www.secureworld.io/industry-news/ships-at-sea-more-ways-to-hack-them.

Simola, J., Pöyhönen, J. and Lehto, M. (2023) "Smart Terminal System of Systems' Cyber Threat Impact Evaluation," *European Conference on Cyber Warfare and Security*, 22(1), pp. 439–449. Available at: https://doi.org/10.34190/eccws.22.1.1070.

SupplyChainBrain (2024) "Chinese Hackers Targeting Ships Across Europe With Malware on USB Sticks," *SupplyChainBrain* [Preprint]. Available at: https://www.supplychainbrain.com/articles/40018-chinese-hackers-targeting-ships-across-europe-with-malware-on-usb-sticks.

Tam, K., Jones, K. and Papadaki, M. (2016) "Threats and Impacts in Maritime Cyber Security," *Engineering & Technology Reference*, 1. Available at: https://doi.org/10.1049/etr.2015.0123.

Tam, K. and Jones, K.D. (2018) "Maritime cybersecurity policy: the scope and impact of evolving technology on international shipping," *Journal of Cyber Policy*, 3(2), pp. 147–164. Available at: https://doi.org/10.1080/23738871.2018.1513053.

The Maritime Executive (2024) "Chinese Spy Malware Found in European Shipping Companies' Systems." Available at: https://maritime-executive.com/article/chinese-spy-malware-found-in-european-shipping-companies-systems.

United States Department of Justice (2015) *Northern District of Oklahoma | Former U.S. Navy Nuclear Systems Administrator Sentenced To 2 Years For Hacking The U.S. Navy And National Geospatial-Intelligence Agency Computer Systems | United States Department of Justice*. Available at: https://www.justice.gov/usao-ndok/pr/former-us-navy-nuclear-systems-administrator-sentenced-2-years-hacking-us-navy-and (Accessed: October 1, 2025).

U.S. Attorney's Office, Northern District of Oklahoma (2014) *United States of America v. Nicholas Paul Knight: Indictment*. Indictment Case No. 4:14-cr-00074-JHP. Tulsa, OK: U.S. District Court for the Northern District of Oklahoma. Available at: https://cdn.theatlantic.com/assets/media/img/posts/050514_Knight.pdf.

ZDNet (2018) "Ships infected with ransomware, USB malware, worms." Available at: https://www.zdnet.com/article/ships-infected-with-ransomware-usb-malware-worms/.

Zetter, K. (2014) “Network Admin Allegedly Hacked Navy—While on an Aircraft Carrier,” *WIRED* [Preprint]. Available at: https://www.wired.com/2014/05/navy-sysadmin-hacking/.

A Governance Model for Cyber Threat Information Sharing in the Healthcare Sector

Jyri Rajamäki[1][1] and Ilkka Tikanmäki[1,2][2]
[1]Business, Data Processing and Service Sector Unit, Laurea University of Applied Sciences, Espoo, Finland
[2]National Defence University, Helsinki, Finland

jyri.rajamaki@laurea.fi
Ilkka.tikanmaki@laurea.fi

Abstract: Cyber Threat Information (CTI) sharing is a vital component of cybersecurity in the healthcare sector, where protecting sensitive patient data and the continuity of critical services are paramount. It's implementing faces socio-technical complexity and strict EU requirements, including the General Data Protection Regulation (GDPR), the Network and Information Security Directive 2 (NIS2), the Cyber Resilience Act (CRA), and the AI Act. This paper applies the Design Science Research (DSR) methodology to develop a governance model that enables secure, compliant, and context-aware CTI sharing. The model integrates systems theory, socio-technical principles, information systems governance, and cyber resilience. It is informed by empirical studies and practical insights from SOC/CERT frameworks. It leverages the DYNAMO platform and tools such as Early Warning System (EWS), open-source software solution MISP, and Data Anonymisation Tool (DAT), to support structured, interoperable, and regulation-compliant threat intelligence exchange. A phased implementation strategy is outlined, beginning with pilot testing in hospitals, then regional integration with CERTs, culminating in national deployment. Evaluation is conducted using realistic assessment and case analysis, with metrics guiding iterative refinement. The model addresses the research question: *How can a governance model for cyber threat information sharing be designed for the healthcare sector under EU regulatory constraints?* This work contributes a scalable, adaptable governance framework that enhances cyber resilience and fosters trust-based collaboration across healthcare ecosystems.

Keywords: Cyber threat intelligence, Healthcare cybersecurity, ECHO early warning system, Interoperability, Design science

1. Introduction

Cybersecurity in healthcare is a growing concern due to the sensitivity of patient data and the critical nature of healthcare services. Recent attacks have demonstrated the vulnerability of hospital systems and the lack of coordinated response mechanisms. The healthcare sector faces increasing cybersecurity threats, ranging from ransomware attacks to data breaches involving sensitive patient information. Sharing cyber threat intelligence (CTI) among stakeholders, such as hospitals, national CERTs, and EU-level agencies, is essential for proactive defence. However, CTI in healthcare is constrained by legal, technical, and organisational factors. Existing CTI sharing systems are either fragmented or nonexistent. The European network of Cybersecurity centres and competence Hub for innovation and Operations (ECHO) project introduced the ECHO Early Warning System (E-EWS) to facilitate cross-sectoral threat sharing, while the Malware Information Sharing Platform (MISP) provides a structured taxonomy for external data exchange. (Chalkias et al., 2023)

The Early Warning System plays a pivotal role in the CTI workflow by serving as a central platform for collecting, analysing, and disseminating threat information. Within the Dynamic Resilience Assessment Method, including a combined Business Continuity Management and Cyber Threat Intelligence solution for Critical Sectors (DYNAMO) framework (DYNAMO project, 2024b), E-EWS integrates both internal and external data sources to generate actionable intelligence (Lahdenperä et al., 2022). Internal sources include logs from the Intrusion Detection System/ Intrusion Prevention System (IDS/IPS), honeypots, and organisational infrastructure, while external sources span Surface Web, Deep Web, and Dark Web content.

Gathering threat data, the system employs multiple crawling techniques. Simple Crawling for indexing URLs, Focused Crawling for CTI-relevant content based on confidence thresholds, and Evasive Crawling to mimic human browsing behaviour and avoid bot detection. These methods ensure comprehensive coverage of threat vectors and indicators. Once data is collected, it undergoes correlation analysis using rule-based and machine learning techniques. The MISP correlation engine identifies relationships between attributes such as hash values and IP addresses. Advanced methods apply text similarity algorithms to detect related threats across different sources. This enables the mapping of tactics, techniques, and procedures (TTPs) used by threat actors (Sun et

[1]https://orcid.org/0000–0003–4798–2462

[2]https://orcid.org/0000- 0001-8950-5221

al., 2023). To ensure privacy and compliance, the Data Anonymisation Tool (DAT) is used to sanitise CTI reports before sharing. It applies Natural Language Processing (NLP) to extract sensitive entities and anonymise them according to predefined policies, protecting identifiers, quasi-identifiers, and confidential attributes (DYNAMO project, 2024a).

Structured sharing of CTI is facilitated through the MISP and Structured Threat Information Expression (STIX) 2.1 format, enabling interoperability and secure exchange of threat intelligence across organisations. Fine-grained access control is enforced using Attribute-Based Encryption (ABE), ensuring that only authorised entities can decrypt and access shared data. This integrated approach enhances situational awareness, supports proactive defence strategies, and strengthens cyber resilience in the healthcare sector. (DYNAMO project, 2024a)

This paper addresses the research problem: How can a governance model for cyber threat information sharing be designed for the healthcare sector under EU regulatory constraints?

The rest of this article is structured as follows: Section 2 presents the Design Science Research methodology; Section 3 describes the healthcare cybersecurity environment and regulatory context. Section 4 presents the knowledge base underlying the governance model. Section 5 provides a detailed explanation of the governance model design. Section 6 focuses on the evaluation and implementation strategy. Finally, Section 7 summarises the article and provides future research directions.

2. Methodology: Design Science Research Framework

The study follows Hevner's Design Science Research (DSR) framework, as shown in Fig. 1 (Hevner, 2007). It consists of three interrelated cycles: relevance, rigour and design. The practical problem is defined in the relevance cycle; the need for a governance model enables a secure and legally compliant Clinical Trials Information System (CTIS) in healthcare. In the rigour cycle, the design is grounded in existing knowledge bases, including systems theory, socio-technical theory, information systems governance models, and risk management theories. Iterative development and evaluation of the governance model is done during the design cycle using design models, evaluation frameworks (e.g. realistic assessment, case analysis), and modeling methods like Business Process Model and Notation (BPMN) and Unified Modeling Language (UML).

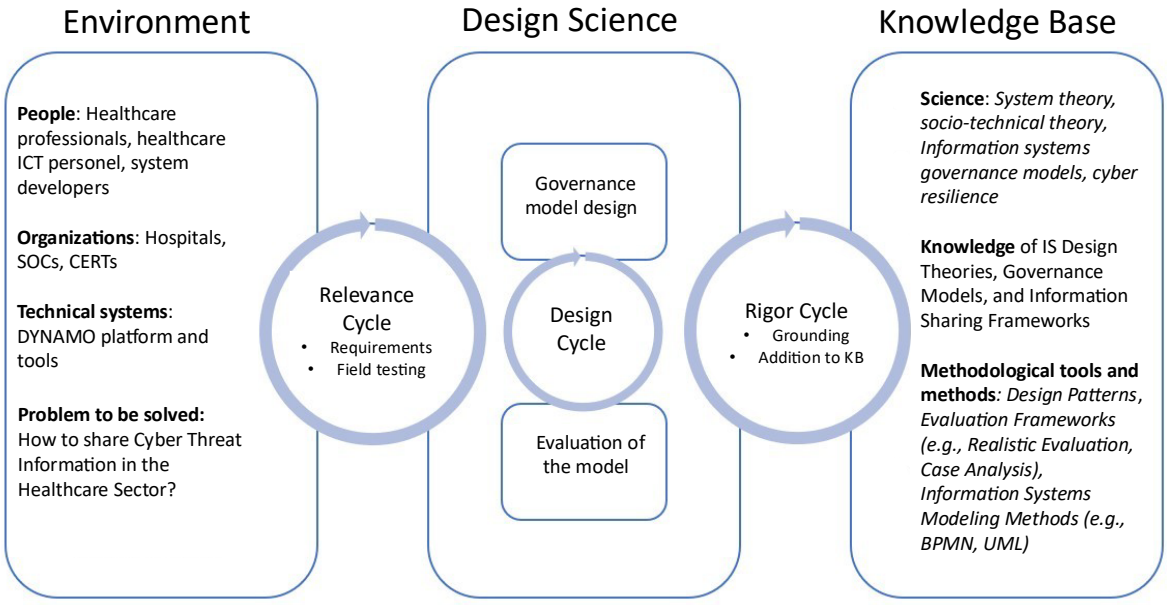


Figure 1: Applied Design Science Research Framework. Adapted from (Hevner, 2007)

Sources include the World Health Organization (WHO), the European Union Agency for Cybersecurity (ENISA), EU policy documents, and peer-reviewed literature. A model-based analysis evaluates the applicability of ECHO's governance model and MISP integration, focusing on role-based access control, interoperability, and legal compliance.

3. Environment

3.1 People

Healthcare professionals, ICT personnel, and system developers constitute the core stakeholder groups in this research. Each of them is contributing essential expertise to the design and implementation of cyber threat information sharing systems in the healthcare sector. Healthcare professionals provide domain-specific knowledge and clinical insights that are critical for aligning technological solutions with real-world medical workflows and patient care requirements. Their involvement ensures that the governance model and supporting systems are not only technically sound but also practically applicable and user-centred.

Healthcare ICT personnel serve as intermediaries between clinical operations and technological infrastructure. Their dual expertise in healthcare processes and information systems enables them to facilitate the integration, configuration, and maintenance of cybersecurity tools and platforms within healthcare organisations. They play a pivotal role in operationalising threat intelligence workflows and ensuring compliance with institutional and regulatory standards.

System developers contribute the technical foundation necessary for building secure, scalable, and interoperable systems. Their collaboration with healthcare professionals and ICT personnel is essential for translating functional requirements into robust software architectures. Developers are responsible for implementing features such as data anonymisation, threat correlation engines, and secure sharing protocols, all of which are central to the proposed governance model. Together, these groups form a multidisciplinary ecosystem that supports the development of resilient and regulation-compliant cyber threat information sharing frameworks tailored to the healthcare domain.

3.2 The DYNAMO Platform and Toolset

The DYNAMO platform is an AI-driven framework developed under the Horizon Europe programme to enhance cyber resilience across critical sectors, including healthcare, energy, and maritime domains. Its primary objective is to support decision-making in cyber threat mitigation and business continuity management (BCM) through integrated threat intelligence workflows. In the healthcare context, DYNAMO facilitates secure, regulation-compliant, and context-aware cyber threat information (CTI) sharing among stakeholders.

At the core of DYNAMO is the Early Warning System (EWS), which serves as a central hub for collecting, analysing, and disseminating threat intelligence. EWS aggregates data from both internal and external sources. Internal sources include intrusion detection and prevention system (IDS/IPS) logs, honeypots, and organisational infrastructure telemetry.

External sources span the Surface Web, Deep Web, and Dark Web, enabling comprehensive threat landscape coverage. EWS gathers threat intelligence through a multi-layered indexing architecture. Simple Crawling indexes URLs for general threat indicators, Focused Crawling targets CTI-relevant content based on trust thresholds, and Evasive Crawling mimics human browsing behaviour to bypass bot detection mechanisms.

Once data is acquired, it undergoes correlation analysis using both rule-based logic and machine learning algorithms. The platform leverages the MISP correlation engine to identify relationships between threat attributes such as hash values, IP addresses, and domain names. Advanced text similarity techniques are applied to detect related threats across heterogeneous sources, enabling the mapping of tactics, techniques, and procedures (TTPs) employed by threat actors.

To ensure compliance with privacy regulations such as GDPR, the Data Anonymisation Tool (DAT) is integrated into the workflow. DAT utilises Natural Language Processing (NLP) to identify and anonymise sensitive entities in CTI reports, including identifiers, quasi-identifiers, and confidential attributes, based on predefined policy rules. (DYNAMO project, 2024a)

Structured sharing of CTI is facilitated through the Malware Information Sharing Platform (MISP) and the STIX 2.1 format, ensuring interoperability and semantic consistency across organisations. Fine-grained access control is enforced using Attribute-Based Encryption (ABE), which guarantees that only authorised entities can decrypt and access shared intelligence.

This integrated toolset enhances situational awareness, supports proactive defence strategies, and strengthens cyber resilience in healthcare environments. It also aligns with EU regulatory frameworks, making it a viable foundation for governance models aimed at secure and compliant CTIS implementation. (DYNAMO project, 2024a)

3.3 Regulatory Context and its Impact on CTIS

EU regulations are significantly shaping the design space for CTIS, with GDPR setting strict rules for sharing personal data, requiring anonymisation and legal justification. NIS2 requires breach reporting and cooperation between key actors, including healthcare providers. The CRA introduces cybersecurity requirements for connected devices, which are relevant for medical IoT systems.

The AI Act regulates the use of AI in sensitive sectors, such as healthcare, and impacts automated threat detection and response. These provisions create both constraints and opportunities for governance design, as highlighted in recent studies and policy documents.

3.4 Compliance within the EU Cybersecurity Legal Framework

The EU cybersecurity regulatory landscape has evolved significantly in recent years, particularly with the introduction of the NIS2 Directive, the Cyber Resilience Act (CRA), and the Corporate Sustainability Reporting Directive (CSRD). This article examines how external Security Operations Centre (SOC) services can support organisations in demonstrating compliance with these regulations. SOC services can substantially enhance an organisation's ability to meet regulatory requirements, produce audit-ready documentation, and report cybersecurity measures to stakeholders.

EU cybersecurity legislation requires organisations to implement technical, organisational, and operational measures to manage cyber risks. Demonstrating compliance is a critical aspect of regulatory implementation, and external SOC services can play a key role in this process. SOCs provide continuous threat monitoring, incident response, log management, and reporting—core elements for evidencing compliance.

3.4.1 NIS2 directive and SOCsServices

The NIS2 Directive (EU 2022/2555) mandates risk-based cybersecurity measures (Article 21) and requires management involvement in cybersecurity governance (Article 20). SOCs can support organisations by:

- Providing real-time threat monitoring and response (Article 21)
- Assisting with incident notification obligations (Article 23): 24h early warning, 72h incident report, 1-month final report
- Supporting executive training and accountability (Article 20)

The directive aims to harmonise cybersecurity practices across the EU and improve the quality and comparability of reporting (Vandezande, 2023).

3.4.2 Cyber resilience act and technical support from SOCs

The CRA (EU Proposal 2022/454) sets requirements for digital products, especially IoT devices (European Commission, 2022). SOCs can assist manufacturers and distributors by:

- Managing and reporting vulnerabilities (Annex I, Section 2)
- Supporting technical documentation and risk assessments (Article 10)
- Ensuring compliance with certification and standards (Article 18)

Compliance with the CRA can be demonstrated through logs, technical reports, and risk management processes—services that SOCs are well-equipped to provide (Shaffique, 2024).

3.4.3 CSRD, ESRS S4, and SOC-generated reporting data

The Corporate Sustainability Reporting Directive (CSRD) (European Parliament, 2022) and European Sustainability Reporting Standard (ESRS S4) require companies to report how they manage risks affecting consumers and end-users, including data breaches. SOCs can contribute by:

- Providing evidence of implemented technical and organisational measures (GDPR Article 32, NIS2 Article 21)
- Supplying audit-ready documentation (logs, threat reports, incident records)
- Supporting management in risk governance and reporting (NIS2 Article 20)

Boggini (2024) demonstrates that compliance with GDPR and NIS2 directly enhances CSRD-aligned reporting and stakeholder transparency. External SOC services can significantly support organisations in meeting the requirements of EU cybersecurity legislation. SOCs contribute to:

- Fulfilling technical and organisational obligations (NIS2, CRA)
- Enhancing audit readiness and reporting quality (CSRD, ESRS)
- Supporting executive accountability and training (NIS2 Article 20)

Integrating SOC services into an organisation's cybersecurity strategy is an effective way to demonstrate compliance and build stakeholder trust.

4. Knowledge Base

This section provides a critical synthesis of existing literature. It combines methodologies, highlights research gaps, and evaluates relevance to the proposed governance model. The governance model proposed in this research is grounded in a robust and multidisciplinary knowledge base that integrates empirical evidence, scientific theory, and methodological rigour.

4.1 Empirical Foundations

The model is based on a selection of key empirical studies examining cybersecurity vulnerabilities, threat landscapes, and operational realities in the healthcare sector. Ewoh and Vartiainen (2024) provide a socio-technical overview of systemic vulnerabilities in healthcare environments, emphasising the interplay between human factors, legacy systems, and organisational complexity.

Vilakazi and Adebesin (2023) present a systematic analysis of cybersecurity threats and mitigation strategies, emphasising the prevalence of ransomware, phishing, as well as the importance of encryption and anomaly detection. Alzaabi and Mehmood (2024) present a comprehensive analysis of insider threat detection, emphasising the limitations of classical machine learning approaches and highlighting the promise of deep learning and NLP-based methods, particularly those leveraging large language models, for enhancing detection capabilities through contextual understanding and temporal pattern analysis.

Aijaz et al. (2023) evaluate threat modeling methodologies in the healthcare IT sector and provide insights into the applicability of structured methodologies for identifying and assessing cyber risks. Veriti (2025) report and the Health-ISAC annual threat report (Health-ISAC, 2025) provide up-to-date information on threat actor behaviour, attack vectors, and sector-specific vulnerabilities, highlighting the need for coordinated CTIS practices. These studies collectively underscore the necessity of integrating human, technical, and organisational dimensions into the design of CTIS governance frameworks.

4.1.1 Comparative critique

Ewoh and Vartiainen (2024) highlight socio-technical vulnerabilities in their article, emphasising human factors and organisational complexity. Vilakazi and Adebesin (2023) focus on technical migration strategies such as encryption and anomaly detection. In contrast, Alzaabi and Mehmood (2024) advocate advanced insider threat detection using deep learning and NLP.

Aljaabi's and Mehmood's article differs from Aijaz et al's argument, which prioritises structured threat modeling methods. These variations highlight the research tension between a proactive governance framework and reactive technical controls. There is also a gap in integrating governance models with adaptive technical solutions under strict EU regulatory constraints, and this study aims to address this.

4.2 Scientific Foundations

The theoretical foundations of the governance model emerge from several established scientific domains. Systems Theory provides a holistic perspective for understanding the dependencies between technological components, organisational processes, and external environments(Becker et al., 2023). Socio-Technical Theory emphasises the co-evolution of social structures and technical systems and advocates for design that balances usability, security, and institutional constraints (Malatji et al., 2019).

Information Systems Governance Models provide structured approaches for aligning IT operations with strategic goals, regulatory requirements, and risk management practices. Cyber Resilience Theory in healthcare describes a model's ability to anticipate, withstand, and recover from cyber disruptions, particularly in critical healthcare infrastructures (Salami, 2025).

4.3 Design Science and Methodological Tools

The model development process is guided by principles of Information Systems Design Theory, which supports the creation of purposeful, context-sensitive artefacts. It also incorporates:

- Governance Models that define roles, responsibilities, and decision-making structures for CTIS.
- Information Sharing Frameworks that facilitate secure, interoperable, and policy-compliant data exchange across organisational boundaries.

To ensure methodological rigour, the following tools and techniques are employed:

- Design Patterns that encapsulate reusable solutions to common design challenges in cybersecurity and information sharing.
- Evaluation Frameworks, including Realistic Evaluation and Case Analysis, which enable context-sensitive assessment of artefact effectiveness and stakeholder relevance.
- Information Systems Modeling Methods, such as Business Process Model and Notation (BPMN) and Unified Modeling Language (UML), which support the formal representation of workflows, system interactions, and governance structures.

Together, these elements form a comprehensive knowledge base that supports the design, implementation, and evaluation of a governance model for cyber threat information sharing in the healthcare sector under EU regulatory constraints.

5. Governance Model Design

The proposed governance model for cyber threat information sharing (CTIS) in the healthcare sector is designed to align technical capabilities with organisational processes and regulatory obligations. It builds upon the DYNAMO platform's architecture and integrates sector-specific governance adaptations, stakeholder roles, and evaluation mechanisms to ensure secure, compliant, and effective CTI exchange.

5.1 Roles and Responsibilities

The governance model defines clear roles for key actors involved in CTIS. The Security Operations Centre (SOC) and Computer Emergency Response Teams (CERT) are responsible for threat detection, response, and coordination across organisational boundaries. Data Protection Officers (DPOs) ensure compliance with data protection regulations such as GDPR and the Health Insurance Portability and Accountability Act (HIPAA).

Healthcare administrators oversee operational integration and policy implementation in clinical settings. Platform administrators, sector-specific administrative directors, and compliance officers (as defined in the DYNAMO pilot) support technical oversight, regulatory alignment, and ongoing monitoring. These roles are structured to support a phased implementation, including preparation, execution, and refinement, as outlined in the DYNAMO pilot strategy.

5.2 Information Sharing Processes

The governance model includes structured processes for secure and interoperable CTI data exchange. Secure communication channels are created using standardised protocols such as STIX/TAXII. Anonymisation protocols are monitored using tools such as the Data Anonymisation Tool (DAT), which applies NLP techniques to sanitise sensitive data. Escalation procedures are defined to ensure that detected threats are responded to quickly, and there are clear coordination paths between stakeholders. These processes are embedded in the DYNAMO platform's Early Warning System (EWS), which aggregates internal and external threat data and supports automated dissemination.

5.3 Compliance Mechanisms

To ensure regulatory alignment, the governance model integrates:

- Embedded compliance checks for GDPR, NIS2, CRA, and the AI Act.
- Policy frameworks that define trust groups, data handling rules, and sector-specific governance adaptations (e.g., HIPAA in healthcare, NERC-CIP in energy).
- Evaluation metrics such as CTI sharing effectiveness, governance adoption rate, and stakeholder trust index, which guide iterative refinement.

These mechanisms are validated through pilot testing, as described in the DYNAMO implementation framework.

5.4 Technology Integration

The governance model leverages the technical capabilities of the DYNAMO platform, including:

- Early Warning System (EWS) for threat detection and alerting.

- Correlation Engine for identifying relationships between threat indicators using rule-based and machine learning techniques.
- MISP and STIX 2.1 for structured data exchange and interoperability.
- Attribute-Based Encryption (ABE) for fine-grained access control.

To support modeling and implementation, the governance framework employs:

- Business Process Model and Notation (BPMN) for visualising workflows.
- Unified Modeling Language (UML) for system architecture and role interactions.
- Design Patterns and Evaluation Frameworks (e.g., Realistic Evaluation, Case Analysis) to guide artefact development and assessment.

Together, these components ensure that the governance model is not only theoretically sound but also practically viable, adaptable to sector-specific needs, and scalable across critical infrastructure domains.

6. Evaluation and Implementation Strategy

The governance model will be evaluated using realistic evaluation and case analysis in selected healthcare organisations. These methods allow for context-sensitive assessment of how well the governance principles translate into operational effectiveness, stakeholder trust, and regulatory compliance. Evaluation metrics such as CTI sharing effectiveness, governance adoption rate, and stakeholder trust index will guide iterative refinement.

To ensure scalability and sustainability, a phased rollout strategy is proposed:

1. Pilot testing in selected hospitals, where the governance model and DYNAMO tools are deployed in controlled environments to validate technical functionality and compliance mechanisms.
2. Regional integration with CERTs, enabling coordinated threat response and intelligence sharing across healthcare institutions and regional cybersecurity teams.
3. National deployment coordinated by health authorities, ensuring alignment with national cybersecurity strategies, regulatory frameworks, and public health priorities.

This staged approach supports continuous stakeholder engagement, empirical feedback collection, and adaptive governance refinement. These allow the model to evolve in response to sector-specific challenges and operational realities.

7. Conclusion

This study demonstrates that applying a Design Science Research (DSR) methodology provides a robust foundation for developing a governance model for cyber threat information (CTI) sharing in the healthcare sector. By integrating theoretical insights from systems theory, socio-technical theory, information systems governance, and cyber resilience, alongside empirical evidence and practical implementation frameworks, the proposed model addresses the multifaceted challenges of CTIS under EU regulatory constraints. The governance model is not only theoretically grounded but also operationally viable. It aligns with the technical capabilities of platforms such as DYNAMO, incorporates structured processes for secure and compliant information exchange. It defines clear roles and responsibilities for key stakeholders, including SOCs, CERTs, healthcare professionals, and ICT personnel. The model also embeds compliance mechanisms for GDPR, NIS2, CRA, and the AI Act, ensuring legal alignment across jurisdictions. A phased implementation strategy, beginning with pilot testing in hospitals, followed by regional integration with CERTs, and culminating in national deployment, supports scalability, stakeholder engagement, and iterative refinement. Evaluation through realistic assessment and case analysis ensures that the governance principles are validated in real-world contexts. Looking ahead, future research should focus on extending the model to support cross-border CTIS, integrating AI-assisted threat detection capabilities, and enabling dynamic compliance monitoring. These directions will further enhance the adaptability, resilience, and regulatory robustness of CTIS governance in healthcare and other critical infrastructure sectors. Ultimately, this work lays the groundwork for a scalable, secure, and regulation-compliant governance framework that strengthens cyber resilience and fosters trust-based collaboration across the healthcare ecosystem.

Acknowledgements

Acknowledgement is paid to the DYNAMO Project, funded by the European Union under grant agreement no. 101069601. However, views and opinions expressed are those of the authors only and do not necessarily reflect

those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Ethics declaration: Ethical clearance was not required for the research.

AI declaration: Keenious Plus was used to assist with literature searching in this work. Copilot was used only in a limited manner to assist with paraphrasing and idea generation during the writing process. No AI tools were involved in producing the core analysis, arguments, or conclusions. The author takes full responsibility for the originality, accuracy, and final content of the paper.

References

Aijaz, M., Nazir, M., & Mohammad, M. N. A. (2023). Threat Modeling and Assessment Methods in the Healthcare-IT System: A Critical Review and Systematic Evaluation. *SN Computer Science*, *4*(6), 714. https://doi.org/10.1007/s42979-023-02221-1

Alzaabi, F. R., & Mehmood, A. (2024). A Review of Recent Advances, Challenges, and Opportunities in Malicious Insider Threat Detection Using Machine Learning Methods. *IEEE Access*, *12*, 30907–30927. https://doi.org/10.1109/ACCESS.2024.3369906

Becker, A. H., Goode, C. H., Rivers, J. C., Tyler, M. W., & Becker, J. J. D. (2023). Shared Governance and Systems Theory: A Mixed Methods Study of Faculty Perceptions and Ideas. *Higher Education Politics & Economics*, *9*(2), 22–47. https://doi.org/10.32674/hepe.v9i2.5974

Boggini, C. (2024). Reporting cybersecurity to stakeholders: A review of CSRD and the EU cyber legal framework. *Computer Law & Security Review*, *53*, 105987. https://doi.org/10.1016/j.clsr.2024.105987

Chalkias, I., Yucel, C., Mallis, D., Rajamäki, J., De Vecchis, F., Hagstrom, P., & Katos, V. (2023). An Empirical Evaluation of Cyber Threat Intelligence Sharing in the ECHO Early Warning System. *Communications in Computer and Information Science Series [In Press]*, *1790*.

DYNAMO project. (2024a). *Initial prototypes of the cyber-threat intelligence gathering, extraction, sharing components and AI-based solutions* (Deliverable No. D4.1; p. 73). https://horizon-dynamo.eu/wp-content/uploads/2025/02/DYNAMO-RPT-D41-V2-0.pdf

DYNAMO project. (2024b, May 9). *Dynamic Resilience Assessment Method Including Combined Business Continuity Management and Cyber Threat Intelligence solution for Critical Sectors*. https://horizon-dynamo.eu/wp-content/uploads/2023/01/DYNAMO_Leaflet_web.pdf

European Commission. (2022). *Cyber Regulation Act (CRA)* (Regulation No. COM(2022) 454 final). European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0454

European Parliament. (2022). *Corporate Sustainability Reporting Directive* (Directive No. 2022/2464; p. 66). European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2464

Ewoh, P., & Vartiainen, T. (2024). Vulnerability to Cyberattacks and Sociotechnical Solutions for Health Care Systems: Systematic Review. *Journal of Medical Internet Research*, *26*, e46904. https://doi.org/10.2196/46904

Health-ISAC. (2025). *2025 Health Sector Cyber Threat Landscape* (Health Sector Cyber Threat Landscape No. 2025; p. 27). Health-ISAC, Inc. https://health-isac.org/wp-content/uploads/Health-ISAC_2025-Annual-Threat-Report.pdf

Hevner, A. R. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Educational Research*, *19*(2), 1–6.

Lahdenperä, J., Muhonen, J., & Rajamäki, J. (2022). How to Utilize E-EWS as a Tool in Healthcare. *European Conference on Cyber Warfare and Security*, *21*(1), 438–441. https://doi.org/10.34190/eccws.21.1.401

Malatji, M., Von Solms, S., & Marnewick, A. (2019). Socio-technical systems cybersecurity framework. *Information and Computer Security*, *27*(2), 233–272. https://doi.org/10.1108/ICS-03-2018-0031

Salami, I. A. (2025). Modeling and Measuring the Cyber Resilience of Critical Healthcare Infrastructure against Ransomware: A Cyber-Physical Systems Risk Perspective. *Journal of Engineering Research and Reports*, *27*(5), 231–252. https://doi.org/10.9734/jerr/2025/v27i51504

Shaffique, M. R. (2024). Cyber Resilience Act 2022: A silver bullet for cybersecurity of IoT devices or a shot in the dark? *Computer Law & Security Review*, *54*(September 2024), 106009. https://doi.org/10.1016/j.clsr.2024.106009

Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., & Zhang, J. (2023). Cyber Threat Intelligence Mining for Proactive Cybersecurity Defense: A Survey and New Perspectives. *IEEE Communications Surveys & Tutorials*, *25*(3), 1748–1774. https://doi.org/10.1109/COMST.2023.3273282

Vandezande, N. (2023). Cybersecurity in the EU: How the Nis2-Directive Stacks Up Against its Predecessor. *Computer Law & Security Review*, *52*, 105890. http://dx.doi.org/10.2139/ssrn.4383118

Veriti. (2025). *The State of Healthcare Cybersecurity 2025* (p. 18) [Research Report]. VERIT, a Check Point Company. https://veriti.ai/wp-content/uploads/2024/12/The-State-of-Healthcare-Cybersecurity-2025-_-A-Veriti-Research-Report.pdf.

Vilakazi, K., & Adebesin, F. (2023). A Systematic Literature Review on Cybersecurity Threats to Healthcare Data and Mitigation Strategies. In A. Gerbera & K. Hinkelmann (Eds.), *EPiC Series in Computing* (Vol. 93, pp. 240–251). EasyChair. https://doi.org/10.29007/hf15

Designing Modular Cybersecurity Education for Healthcare: Integrating XiA and ManagiDiTH Frameworks with CyberSecPro Insights

Jyri Rajamäki[1] and Jussi Simola[2]
[1]Laurea University of Applied Sciences, Espoo, Finland
[2]University of Jyväskylä, Finland

jyri.rajamaki@laurea.fi
simolajussi@gmail.com

Abstract: The healthcare sector faces increasingly complex cybersecurity threats due to its rapid digital transformation, including the adoption of electronic health records (EHRs), connected medical devices, and telehealth services. This paper presents a modular and interdisciplinary model for cybersecurity education tailored to healthcare professionals. The model integrates three EU-funded initiatives: the XiA Framework's microlearning approach, the ManagiDiTH master's programme's strategic and human-centric curriculum, and the CyberSecPro project's hands-on, scenario-based training modules. Together, these components form a comprehensive educational ecosystem that supports continuous professional development, microcredentialing, and alignment with European standards such as the European Qualifications Framework (EQF), the NIST Cybersecurity Framework, and ENISA's European Cybersecurity Skills Framework (ECSF). The paper outlines the pedagogical foundations of microlearning, the structure of Micro-Content Learning Blocks (MCLBs), and the curriculum design principles that foster cyber resilience in healthcare. It concludes with a discussion on implementation potential, impact indicators, and future directions for research and practice.

Keywords: Cybersecurity education, Healthcare cybersecurity, Microlearning, Microcredentials, European cybersecurity skills framework (ECSF)

1. Introduction

The digital transformation of healthcare has brought significant advancements in patient care, diagnostics, and operational efficiency. However, this evolution has also introduced complex cybersecurity challenges that threaten the integrity, confidentiality, and availability of health data. Hospitals and healthcare organizations are increasingly targeted by cyberattacks, including phishing, ransomware, and Distributed Denial of Service (DDoS) attacks, which can disrupt critical services and compromise sensitive patient information (Priestman *et al.*, 2019; Zhou *et al.*, 2024).

Despite growing awareness, traditional security measures remain insufficient to address the evolving threat landscape. Artificial intelligence (AI) offers promising capabilities for enhancing cybersecurity in healthcare by enabling proactive threat detection and adaptive response mechanisms. AI-driven tools, such as SecureAI, demonstrate how learning from historical patterns can help identify anomalous network behaviors and mitigate risks in real-time (Rajamäki, Nepal and Chalkias, 2025). Healthcare organizations, particularly private entities, often lag behind other sectors in implementing robust cybersecurity frameworks (Tavares, Sousa and Proença, 2024). The urgency to adopt AI-based solutions is underscored by the increasing complexity of digital health infrastructures and the growing reliance on interconnected systems, such as the Internet of Medical Things (IoMT). Aljanabi (2023) illustrates how trustworthy AI techniques can harden intrusion detection systems against data poisoning threats, ensuring continuity of care even under hostile conditions. However, the integration of AI into healthcare cybersecurity is not without ethical and regulatory concerns. Issues such as algorithmic bias, lack of transparency, and accountability must be addressed through comprehensive governance frameworks. Zhang and Zhang (2023) advocate for ethical and legal mechanisms that align with the EU AI Act, which classifies medical AI applications as high-risk and mandates stringent standards for safety and reliability. Moreover, safeguarding patient data requires advanced privacy-preserving techniques. Approaches such as anonymization, differential privacy, and secure multiparty computation are essential for balancing data utility with confidentiality (Ezeah, 2024; Forood *et al.*, 2024). These methods are particularly critical in domains like cancer care, where AI-based diagnostics and precision medicine rely heavily on sensitive health information.

To address the abovementioned challenges, healthcare professionals must be equipped with specialised cybersecurity competencies. This paper proposes a modular education model that integrates the XiA Framework's microlearning approach (XiA, 2025), the interdisciplinary structure of the ManagiDiTH master's programme (ManagiDiTH, 2025), and the market-driven insights of the CyberSecPro project (CyberSecPro, 2025). The remainder of the paper is structured as follows: Section 2 outlines the research methodology,

including the design-based approach and comparative analysis used to synthesize the three educational frameworks. Section 3 presents an in-depth overview of the XiA, ManagiDiTH, and CyberSecPro initiatives, highlighting their pedagogical strategies, thematic coverage, and alignment with European standards. Section 4 focuses on the development of Micro-Content Learning Blocks (MCLBs) for cybersecurity education, providing a detailed examination of their structure and learning outcomes, and elucidating the pedagogical foundations of microlearning in healthcare cybersecurity with particular emphasis on learner-centered design, EQF compatibility, and the support of continuous professional development. Section 5 offers a discussion, reflecting on the model's implementation potential and future directions for research and practice. Finally, Section 6 concludes the paper.

2. Research Methodology

This study employs a qualitative, design-based research methodology as the foundation for developing a modular cybersecurity education model for the healthcare sector. The approach combines document analysis, framework synthesis, and comparative evaluation across three EU-funded projects: XiA, ManagiDiTH, and CyberSecPro. These are selected because the EU-level need of the standardized interoperability in accordance with cybersecurity requirements. The primary method is document analysis, which involves systematic examination of project deliverables, curriculum frameworks, and academic publications. This method is well-established in educational research for extracting structured insights from policy and pedagogical documents (Bowen, 2009). The analysis focused on the pedagogical foundations of micro-content learning and on curriculum design in healthcare cybersecurity. Of the analyzed projects, two (CyberSecPro and ManagiDiTH) have already published their results in openly accessible deliverables and scientific publications, whereas one project launched this year (XiA), does not yet have such materials available. Therefore, the first two projects were analyzed based on public documents, while the analysis of the third relied on internal project working papers.

A comparative case study approach (Yin, 2018) was used to evaluate the similarities and differences between the three initiatives. This included mapping cybersecurity content across modules, identifying target learner profiles, and assessing pedagogical coherence. The comparison informed the integration strategy proposed in the paper, ensuring that the resulting curriculum model reflects best practices and addresses sector-specific needs. Furthermore, the study applies a design-based research (DBR) methodology (Wang and Hannafin, 2005; Reeves, 2006), which is particularly suitable for developing educational interventions in real-world contexts. DBR allows for iterative refinement of the curriculum model through synthesis of theory and practice, guided by principles of microlearning, human-centric cybersecurity, and continuous professional development.

The proposed model is validated through conceptual alignment with recognized educational standards such as the European Qualifications Framework – EQF (European Parliament, 2017), the Finnish National Framework for Qualifications and Other Competence Modules – FiNQF (EDUFI, 2025), the European Cybersecurity Skills Framework – ECSF (ENISA, 2022), and the NIST Cybersecurity Framework (NIST, 2013). While empirical testing is beyond the scope of this paper, future research may include pilot implementations, learner feedback, and impact assessment using key performance indicators such as skill acquisition, incident response capability, and professional mobility.

3. EU-Level Educational Initiatives in Healthcare Cybersecurity

As the healthcare sector undergoes rapid digital transformation, the need for robust cybersecurity competencies has become increasingly urgent. In response, the European Union has launched several strategic educational initiatives aimed at strengthening cybersecurity skills across healthcare professionals, students, and institutions. This section highlights three key projects that contribute to building a resilient and well-trained workforce in digital health. The CyberSecPro Project aims to advance cybersecurity education, with healthcare being one of its three primary domains. In contrast, ManagiDiTH and XiA are development projects focused on digital health education, within which cybersecurity plays a significant role. There is no similar example where cybersecurity challenges of healthcare sector has taken into account by utilizing EU-level requirements and project frameworks in this way. Laurea university of applied sciences engages in development work in all mentioned projects.

Figure 1 illustrates the ecosystem of healthcare cybersecurity education, highlighting the relationships between the CyberSecPro, ManagiDiTH, and XiA projects.

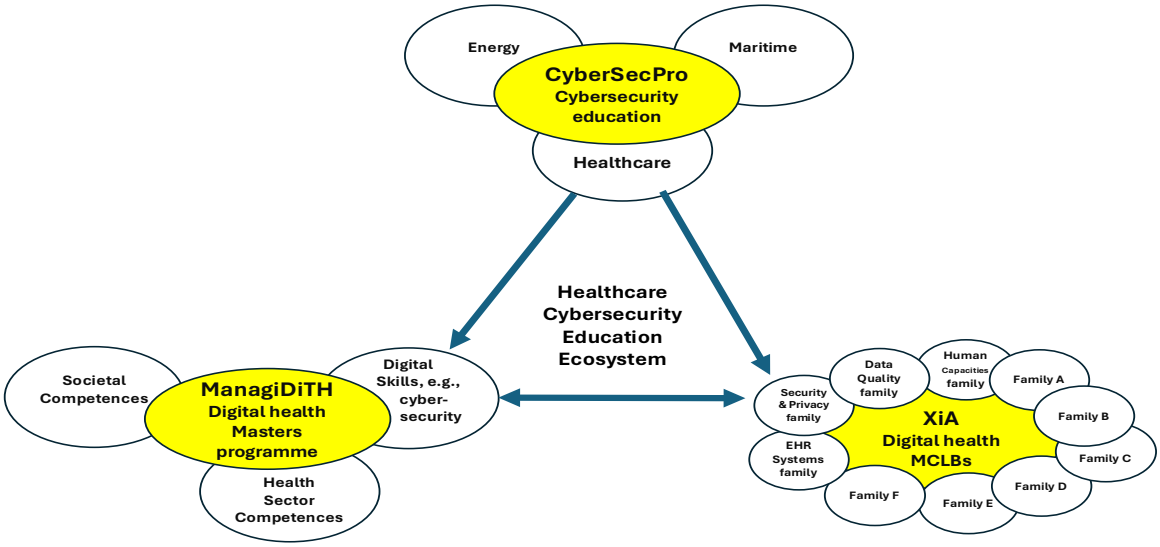


Figure 1: Ecosystem of healthcare cybersecurity education: Relationships between CyberSecPro, ManagiDiTH, and XiA Projects

3.1 CyberSecPro

The CyberSecPro Project is an EU-funded initiative under the Digital Europe Programme, aimed at developing advanced cybersecurity competencies across critical sectors. It covers a wide range of topics, including cybersecurity principles, risk management, security tools, ethical hacking, and incident response. A key focus of the project is human-centric cybersecurity, which recognizes the importance of user behavior, awareness, and ethical considerations in maintaining secure systems. The initiative also promotes continuous professional development, ensuring that professionals remain equipped to handle emerging threats and technologies.

The project's comprehensive market analysis revealed significant gaps in cybersecurity readiness within healthcare, prompting the development of targeted training modules (Rathod *et al.*, 2023). Therefore, a comprehensive portfolio of cybersecurity training modules was tailored to the healthcare sector (Koutras, *et al.*, 2024). This initiative responds to the growing need for healthcare professionals to acquire practical cybersecurity skills to protect sensitive patient data and ensure the integrity of digital health systems. The curriculum is structured into twelve thematic modules, each addressing critical aspects of cybersecurity in healthcare. These modules cover a wide range of topics, including cybersecurity management, human factors, risk governance, network security, data protection, threat intelligence, emerging technologies, infrastructure security, software security, penetration testing, cyber range operations, and digital forensics. Each module is designed to combine both generic cybersecurity principles and healthcare-specific challenges. For example, the module on *Cybersecurity Essentials and Management for Health Sector* introduces foundational concepts such as threat landscapes, regulatory frameworks, and risk mitigation strategies, while also incorporating healthcare-specific case studies and simulations. Similarly, the module on *Human Factors and Cybersecurity for Health* explores psychological and behavioral dimensions of cybersecurity, emphasizing the role of healthcare staff in maintaining cyber hygiene and preventing insider threats (Koutras, *et al.*, 2024).

The training materials emphasize hands-on learning through interactive exercises, simulations, and gamified environments. Modules such as *Cybersecurity Risk Management and Governance* and *Penetration Testing for Healthcare IT Infrastructures* provide learners with practical tools to assess vulnerabilities, implement security controls, and respond to incidents. The inclusion of cyber range environments allows participants to engage in realistic scenarios, simulating attacks and defenses within healthcare IT infrastructures. CyberSecPro also addresses emerging challenges by integrating modules on *Cybersecurity in Emerging Technologies*, including explainable AI and anomaly detection, and *Critical Infrastructure Security*, which examines cascading effects in complex health networks. These modules prepare healthcare professionals to anticipate and respond to future threats in increasingly interconnected and data-driven environments. The curriculum is aligned with European cybersecurity frameworks, including the European Cybersecurity Skills Framework (ECSF), ensuring that the training supports recognized competencies and career pathways. The modular design allows for flexible implementation across academic institutions, professional training programs, and continuing education initiatives (Koutras, *et al.*, 2024).

By offering a holistic and sector-specific approach to cybersecurity education, CyberSecPro empowers healthcare providers to make informed decisions, adopt best practices, and foster a culture of security within their organizations. The project contributes to strengthening the cybersecurity posture of the healthcare sector across Europe, supporting both individual skill development and systemic resilience (Koutras, *et al.*, 2024).

3.2 ManagiDiTH

The ManagiDiTH Programme is a joint master's degree initiative involving universities from three EU member states. It offers a multidisciplinary curriculum that integrates health sector expertise, societal competencies, and digital skills. Within the digital skills track, students may specialize in data science or interoperability, making the programme highly relevant to the evolving requirements of healthcare systems. ManagiDiTH emphasizes cross-border collaboration and mobility, preparing graduates to operate in diverse healthcare environments across Europe (ManagiDiTH, 2025).

The cybersecurity studies within the programme encompass both technical and strategic domains, equipping students with the capabilities to conduct risk assessments, implement protective measures, and participate in the development of digital health governance policies. Cybersecurity is addressed through a dedicated study unit titled *Cybersecurity for Health Systems* (6 ECTS), which provides students with a comprehensive understanding of the unique cybersecurity challenges in healthcare. The course covers the fundamentals of cybersecurity, including risk assessment and management. Students are trained to identify threats, implement safeguards, detect anomalies, respond to incidents, and recover systems. Topics include network and application security, secure cloud computing, medical device protection, governance and compliance, as well as future trends in healthcare cybersecurity. Students also learn to stay informed about emerging cybersecurity trends and technologies, while considering ethical implications and best practices in the healthcare sector. The cybersecurity curriculum is aligned with international standards, such as ECSF and NIST (Rajamäki *et al.*, 2024).

In addition to this core module, cybersecurity themes are embedded across several other study units. For example, *the Managing the Digital Transformation in Healthcare* unit explores ethical and privacy considerations in the use of digital tools. *The Health Data and Information Systems* unit addresses data standards, analytics, and legal frameworks related to health information. *The Regulation, Legislation and Structures in Health* unit examines EU-level digital health governance and its implications for public and private service providers. *The Ethics and Privacy in Health* unit focuses on GDPR compliance and privacy protection. Other units, such as *E-Health and Telemedicine, Interoperable Ecosystems in Health,* and *Emerging Technologies in Interoperability*, also include content related to security and privacy (Rajamäki *et al.*, 2024).

ManagiDiTH adopts a human-centric approach to cybersecurity, recognizing that human behavior is often the weakest link. Based on Grobler's, Gaire's and Nepal's (2021) 3U model (User, Usage, Usability), the curriculum emphasizes situational awareness, psychological and cognitive factors, legal and technical measures, and user experience design. This ensures that students understand not only the technical aspects of cybersecurity but also the human and organizational dimensions. Especially in Internet of Medical Things (IoMT) environments, where numerous interconnected medical devices and systems increase both technical and human vulnerabilities, a comprehensive understanding of user behavior, device usage, and usability becomes particularly important (Rajamäki and Postolache, 2025).

ManagiDiTH emphasizes hands-on learning through case studies and practical exercises, fostering the real-world application of cybersecurity principles. The programme encourages collaboration between educational institutions, healthcare organizations, and cybersecurity experts to ensure relevance and adaptability. Continuous professional development is supported through ongoing training and curriculum updates, and ethical values are reinforced to promote responsible cybersecurity practices. This comprehensive and interdisciplinary approach positions ManagiDiTH as a model for cybersecurity education in healthcare, addressing both current challenges and future needs in a rapidly evolving digital landscape.

3.3 XiA Framework

The XiA framework is a European educational model developed to support the digital transformation of healthcare through modular and flexible learning. It utilizes Micro-Content Learning Blocks (MCLBs), which are short, targeted learning units. These blocks are designed to support micro-credentials, enabling learners to build their competencies in a stackable and customizable manner. The framework is particularly suitable for professionals who require continuous education in rapidly evolving fields such as cybersecurity. Through personalized learning pathways, XiA enables the acquisition of specialized digital health skills without the need

for full degree programs. This approach is ideal for healthcare workers who wish to advance their expertise while remaining active in clinical or administrative roles.

All XiA MCLBs are mapped to EQF levels 5–7, ensuring consistency and recognition across different educational systems. This alignment supports the development of technical and analytical skills (EQF 5–6) for cybersecurity implementation, as well as strategic and leadership competencies (EQF 7) for managing digital transformation in healthcare.

The XiA framework includes a dedicated MCLB family called *Security & Privacy*, which focuses on cybersecurity and data protection in the context of digital health interoperability. This thematic family addresses several critical areas, such as compliance with data protection regulations (e.g., GDPR and HIPAA). It also covers techniques for data anonymization and pseudonymization, implementation of audit trails and access logs, and methods for privacy-preserving cross-border data exchange. Additionally, the content emphasizes building trust between clinicians and patients by ensuring secure and transparent data handling. Within the XiA framework, cybersecurity is regarded as a fundamental requirement for developing reliable and sustainable digital health infrastructures.

The pedagogical design of the XiA framework is based on learner-centeredness, modularity, and competency-based education principles. These principles are essential for addressing the dynamic learning needs of healthcare professionals, especially in the context of cybersecurity, where technological advancements require continuous skill updates. Each XiA MCLB is designed with a clear pedagogical structure that supports effective learning. Modules promote active engagement through simulations, discussions, and problem-solving tasks. Learning progress and competencies are continuously monitored using both formative and summative assessments. Furthermore, multimedia content is integrated to enrich the learning experience and accommodate diverse learning styles.

Cybersecurity is systematically embedded in the competency profiles defined for three key target groups: IT professionals (designers), healthcare and IT management (managers), and healthcare practitioners (end-users). Particular emphasis is placed on competencies related to the responsible handling of health data. These competencies include secure data storage and anonymization techniques, cybersecurity practices such as encryption and access management, knowledge of relevant legislation (GDPR, HIPAA), and mastery of ethical frameworks for the use of artificial intelligence in clinical environments. By integrating these elements, the XiA framework ensures that cybersecurity is addressed not only as a technical skill but also as a core professional responsibility across various roles in healthcare.

Competency-based learning ensures that each block supports practical skills, such as secure data handling, ethical decision-making, and interdisciplinary collaboration. Learning outcomes are recognized through micro-credentials, which validate competencies and enhance employability. These micro-credentials are awarded based on workload and learning outcomes and may be issued by universities or other accredited bodies. They serve as tangible evidence of skills and can be combined into broader qualifications.

In addition to the Security & Privacy family, cybersecurity is addressed in several other MCLB families. For example, the *Data Quality* family focuses on data integrity and its critical role in ensuring the safety and reliability of AI applications in healthcare. The *Electronic Health Record (EHR) Systems* family examines the secure architecture and regulatory aspects of electronic patient record systems. The *Human Capacities* family highlights organizational change management and capacity building to support secure digital transformation. This cross-cutting integration ensures that cybersecurity is not treated solely as a technical issue but is embedded within broader organizational and systemic contexts, reinforcing its significance across multiple domains of digital health education.

To ensure pedagogical integrity, XiA provides a standardized template for MCLB development. The template includes clearly defined learning objectives and corresponding EQF levels, identification of target groups, specification of assessment types and workload, as well as licensing and quality assurance mechanisms. All content is produced under the CC BY-SA 4.0 license, promoting openness, reusability, and co-creation. Pedagogical quality is ensured through peer review, compliance with EU standards, and continuous updates to maintain relevance and currency.

The XiA framework integrates cybersecurity not only as a technical competency but also from ethical and legal perspectives. It addresses, for example, the ethical use of artificial intelligence in clinical environments, emphasizing responsible decision-making and patient safety. MCLBs also promote transparency and privacy in data handling, ensuring that healthcare professionals understand the implications of data use, sharing, and

protection. Additionally, the framework supports compliance with European legislation, such as the AI Act and European Health Data Space (EHDS), highlighting the harmonization of digital health practices with legal standards. Through this multidimensional approach, the XiA framework positions cybersecurity as a cornerstone of responsible innovation and governance in healthcare.

In the context of XiA, Continuing Professional Development (CPD) emphasizes the relevance of professional roles—particularly in healthcare cybersecurity—as well as flexibility and modularity, allowing learners to select blocks that align with their career goals. Practical cybersecurity skills are essential for safeguarding digital healthcare environments. Learners are trained to implement security protocols that protect sensitive health data and ensure system integrity. They also gain expertise in managing audit trails and access control systems, which are critical for monitoring data usage and preventing unauthorized access. The framework includes guidance for the secure integration of healthcare IT systems, such as solutions compliant with the Health Level 7 Fast Healthcare Interoperability Resources (HL7, 2025) standard, to promote interoperability without compromising security. Additionally, learners are introduced to the management of cloud-based healthcare information systems, equipping them to use and protect modern, distributed healthcare infrastructures. These competencies are vital for maintaining the confidentiality, integrity, and availability of health data in increasingly complex and interconnected digital ecosystems.

3.4 Comparative Analysis of Cybersecurity Education Approaches

The XiA Framework, ManagiDiTH programme, and CyberSecPro project each contribute distinct yet complementary approaches to cybersecurity education in the healthcare sector. While all three initiatives aim to strengthen cybersecurity competencies among healthcare professionals, they differ in scope, pedagogical structure, and target audiences.

XiA Framework emphasizes modular microlearning through Micro-Content Learning Blocks (MCLBs), which are aligned with EQF levels 5–7 and support microcredentialing. Cybersecurity is integrated across multiple MCLB families, with a dedicated *Security & Privacy* family addressing GDPR, HIPAA, anonymization, and secure data exchange. XiA's strength lies in its flexibility and adaptability for continuous professional development, making it suitable for both clinical and technical staff seeking targeted upskilling as well as for system developers.

ManagiDiTH, on the other hand, offers a formal joint master's degree programme with a comprehensive curriculum that includes a dedicated 6 ECTS module on *Cybersecurity for Health Systems*. The programme integrates cybersecurity themes across various study units and emphasizes interdisciplinary learning, human-centric design, and alignment with NIST and ECSF frameworks. Its pedagogical model supports strategic thinking and leadership in digital health transformation, targeting professionals with diverse backgrounds in healthcare, ICT, and management.

CyberSecPro provides a sector-specific training portfolio and offers twelve thematic modules for cybersecurity education in the healthcare sector. These modules cover both foundational and advanced topics, including threat intelligence, penetration testing, digital forensics, and emerging technologies. CyberSecPro's approach is highly practical, featuring gamified learning, cyber range simulations, and alignment with ECSF competencies. It is designed for scalable implementation across academic and vocational settings, addressing identified skills gaps through hands-on, scenario-based training.

In summary, XiA provides a flexible microlearning infrastructure, ManagiDiTH offers a structured academic pathway with interdisciplinary depth, and CyberSecPro delivers targeted, practice-oriented training modules. Together, these initiatives form a robust ecosystem for cybersecurity education in healthcare, supporting lifelong learning, professional specialization, and systemic resilience.

4. Developing Micro-Content Learning Blocks for Cybersecurity Education

This section presents an example of how cybersecurity education in the healthcare sector could be approached in the future. The aim is to illustrate a pedagogical and structural model that supports the development of relevant competencies through modular and flexible learning solutions. By outlining practical implementation strategies, this section guides educators and curriculum designers seeking to enhance cybersecurity skills among healthcare professionals in a rapidly evolving digital environment.

4.1 Example Approach for Advancing Cybersecurity Education in Healthcare

The evolving demands of society challenge traditional conceptions of learning and the channels for acquiring knowledge. Individual characteristics and traits must be better acknowledged, and the European Union has

made this perspective a central objective of its learning strategies. As we know, levels of intelligence and understanding of the surrounding world vary among individuals. There is no universal model that works for everyone. Teachers and educational institutions must seize opportunities for flexible learning to respond to environmental changes and remain up to date.

European standardization is systematically monitored in Finland. The National Qualifications Framework (FiNQF) is based on the European Qualifications Framework (EQF). FiNQF describes the qualifications and competence structures of the national education system, covering, for example, general education, vocational education, and higher education. Competence structures are organized into eight levels based on the skills they require. The competencies for each level are defined in a government decree, and the descriptions correspond to EQF requirements. The Finnish Qualifications Framework is aligned with the European Higher Education Area (EHEA) framework (EDUFI, 2025). Curriculum design should be aligned with international standards such as the NIST Cybersecurity Framework (NIST, 2013) and the ECSF European Cybersecurity Skills Framework (ENISA, 2022), ensuring content relevance and transferability across EU member states.

Micro-learning content blocks (MCLBs) represent a modular and flexible approach to professional education, particularly well-suited for rapidly evolving fields such as healthcare cybersecurity. The pedagogical design of MCLBs should emphasize competency-based learning, and each block should begin with clearly defined and measurable objectives developed in accordance with Bloom's taxonomy. The curriculum should be structured in a modular format, enabling the organization of content into micro-courses that can be tailored to the needs of different stakeholders—such as healthcare professionals, IT managers, and system developers.

4.2 An Effective Pedagogical Approach for Cybersecurity Education

A sound pedagogical approach emphasizes experiential learning through case studies, simulations, and cyber range environments. It also incorporates a human-centered perspective, recognizing behavior, usability, and organizational culture as critical dimensions of cybersecurity. This is operationalized through the 3U model (User, Usage, Usability), ensuring that learners develop both technical proficiency and contextual understanding. Engagement and practical application should be fostered through interactive activities such as quizzes, simulations, and collaborative tasks. Both formative and summative assessments should be integrated into the learning experience to monitor progress and reinforce understanding.

Best practice recommends developing MCLBs into thematic families that address broad domains of digital health, such as interoperability, data protection, and cybersecurity. Each MCLB supports a broader learning pathway and can be combined into micro-courses designed to build specific competencies. These competencies are mapped to stakeholder roles, ensuring relevance and applicability in real healthcare settings.

4.3 Micro-credentials as Enablers of Modular Competence Recognition

Learners completing MCLBs or micro-courses may earn micro-credentials, serving as formal recognition of acquired skills. These credentials are aligned with EQF levels and can be issued by universities or other accredited bodies. A typical MCLB corresponds to approximately 0.1–0.2 ECTS credits, depending on workload, and can be aggregated into larger competence units over time.

Micro-credentials support modular recognition of competencies: learners can earn formal acknowledgments for completed modules (typically 0.1–5 ECTS), which can be stacked into broader qualifications, enabling personalized and scalable learning pathways. Micro-credentials enhance professional mobility and visibility, particularly in sectors such as healthcare, where regulatory requirements and technical expertise are critical. Content developers are encouraged to employ diverse formats—text, video, audio, and interactive media—to accommodate different learning styles and accessibility needs. All materials should adhere to open licensing principles to facilitate reuse and collaboration.

The micro-learning approach aligns well with cybersecurity frameworks. The European Cybersecurity Skills Framework (ECSF) supports the identification and structuring of role-based tasks, competencies, skills, and knowledge for European cybersecurity professionals. ECSF is an EU-level framework for defining and assessing competencies, endorsed by the European Commission in 2023 (ENISA 2022). It consolidates cybersecurity roles into 12 profiles, analyzed in terms of responsibilities, skills, synergies, and dependencies. ECSF provides a shared understanding of key roles, competencies, skills, and knowledge most frequently required in cybersecurity. It assists in identifying cybersecurity capabilities and supports the design of cybersecurity education.

Effective pedagogy in micro-learning content begins with clearly defined learning objectives that specify what learners should know or be able to do upon completing a block. Instructional design is grounded in learner-

centeredness, modularity, and competency-based education. These principles are essential for addressing the dynamic learning needs of healthcare professionals, particularly in the context of cybersecurity, where technological advancements demand continuous skill updates. Bloom's taxonomy is applied to formulate learning objectives across cognitive levels—from foundational understanding to advanced synthesis and evaluation. Curriculum design entails aligning learning objectives with teaching methods, assessment strategies, and learning resources. Table 1. illustrates and clarify the pedagogical and structural model supporting modular, competency-based cybersecurity education in healthcare.

The pedagogical framework integrates instructional quality, domain-specific content, and flexible delivery modes, strengthening cyber resilience among healthcare professionals. It promotes lifelong learning and professional mobility, contributing to the development of secure and sustainable digital healthcare systems across Europe.

Table 1: Modular-based Pedagogical and Structural Model for Cybersecurity Education

Model Layer	Key Elements	Pedagogical / Structural Function	XiA Framework Role	ManagiDiTH Master's Program Role	CyberSecPro Module Role
Policy & Standards Alignment	EQF, EHEA, FiNQF, NIST CSF, ENISA ECSF	Ensures international alignment, transferability, and recognition of competencies.	Provides organizational and process-level alignment templates.	Maps program learning outcomes to EQF level 7 and EHEA requirements.	Maps role-based skills and tasks to ECSF and NIST functions.
Pedagogical Principles	Competency-based learning, Bloom's taxonomy, 3U model, experiential learning.	Defines how learning objectives, activities, and assessments are designed.	Supports integration of human, organizational, and technical perspectives.	Implements learner-centered pedagogy and experiential projects.	Emphasizes applied learning through labs and simulations.
Structural Modularity	MCLBs, micro-courses, thematic families, learning pathways	Enables flexible, stackable, and reusable learning units.	Supplies reusable content patterns and case assets.	Curates MCLBs into degree-aligned micro-courses and pathways.	Provides ready-made modular training units.
Learning Design & Delivery	Case studies, cyber ranges, simulations, interactive activities.	Supports practical application and engagement.	Contributes cross-context case libraries.	Coordinates blended delivery and capstone projects.	Delivers role-based labs and scenario exercises.
Assessment Strategy	Formative and summative assessments, authentic tasks.	Measures progress and competency attainment.	Provides evaluation templates and usability metrics.	Defines program-level assessment rubrics.	Uses skills checklists and lab-based assessments.
Credentialing	Micro-credentials, ECTS stacking, digital badges.	Enables formal recognition and lifelong learning pathways.	Supports evidence packaging for micro-credentials.	Issues university-backed micro-credentials and degree credits.	Issues role-aligned micro-credentials.

5. Discussion

This paper has introduced a modular, interdisciplinary, and scalable model for cybersecurity education tailored to the healthcare sector. The model integrates three complementary European initiatives: the XiA Framework, the ManagiDiTH master's programme, and the CyberSecPro training modules. Together, these components form a comprehensive educational ecosystem that addresses the complex and evolving cybersecurity needs of healthcare professionals.

The XiA Framework contributes flexibility and accessibility through its Micro-Content Learning Blocks (MCLBs), which support just-in-time learning and microcredentialing. These blocks are particularly suitable for professionals who require targeted upskilling without committing to full degree programmes. The ManagiDiTH programme adds academic depth and strategic perspective, offering a formal master's degree with interdisciplinary content and a strong emphasis on human-centric cybersecurity. CyberSecPro complements these approaches with hands-on, scenario-based training modules that are aligned with European cybersecurity frameworks and designed to address real-world threats and skills gaps. By combining these strengths, the proposed model supports lifelong learning, professional mobility, and systemic resilience in healthcare cybersecurity. It enables learners to acquire both technical competencies and contextual awareness, fostering a

culture of security across clinical, administrative, and technical roles. The modular nature of the model makes it well-suited for integration into various educational contexts, including continuing professional development (CPD) programmes, vocational training, and higher education. Learners can pursue personalized learning journeys that align with their professional roles and career goals. The model also supports scalable deployment across EU member states, enhancing its relevance and reach.

Formal recognition of learning achievements is facilitated through microcredentials, typically ranging from 1 to 5 ECTS. These credentials validate acquired skills and contribute to employability and career advancement. However, successful implementation of the model requires institutional commitment to curriculum innovation, as well as active collaboration between educational institutions, healthcare providers, and cybersecurity experts.

Challenges may arise in maintaining the currency and relevance of content, especially in a rapidly evolving field like cybersecurity. Therefore, mechanisms for regular curriculum updates, quality assurance, and learner feedback are essential. On the other hand, modularity facilitates the maintenance of up-to-date content. Additionally, ensuring interoperability between different educational systems and credentialing frameworks across Europe will be critical for widespread adoption.

Future research should focus on empirical validation of the model through pilot implementations and longitudinal studies. Learner experience, satisfaction, and outcomes should be systematically evaluated to refine the pedagogical approach and enhance effectiveness. Moreover, expanding the curriculum to include emerging domains such as quantum-safe cryptography, explainable artificial intelligence, and digital ethics will ensure that the model remains responsive to future challenges.

6. Conclusion

In the context of the EU´s digital strategy the frameworks presented and related projects support the needs identified by the EU to harmonize the interconnection of IT skills needs, challenges created by artificial intelligence and education in the healthcare sector where public sector and private sectors actively cooperate. The rapidly evolving AI-assisted digital world of work environment also requires rapid response in education and understanding of professionals. The XIA framework with Micro-Content learning blocks is one example and one way to effectively meet local or regional ethical needs at a practical level, which the healthcare sector, and in particular patient safety, requires when developing common standardization in the sector.

Acknowledgements

This research was conducted within the framework of three EU-funded projects:

ManagiDiTH (Grant No. 101083896)

CyberSecPro (Grant No. 101083594)

XiA – Xpanding Innovative Alliance (Grant No. 101187650)

All projects are supported by the European Health and Digital Executive Agency (HADEA).

Ethics declaration: Ethical clearance was not required for the research.

AI declaration: Artificial intelligence (AI) was utilized to support language refinement, and structural coherence. AI-assisted tools were employed to analyze and integrate information from multiple EU-funded projects, ensuring consistency in terminology and alignment with pedagogical frameworks. While all conceptual contributions and interpretations remain the responsibility of the author.

References

Aljanabi, M. (2023) 'Safeguarding Connected Health: Leveraging Trustworthy AI Techniques to Harden Intrusion Detection Systems Against Data Poisoning Threats in IoMT Environments', *Babylonian Journal of Internet of Things*, 2023, pp. 31–37. Available at: https://doi.org/10.58496/BJIoT/2023/005.

Bowen, G.A. (2009) 'Document Analysis as a Qualitative Research Method', *Qualitative Research Journal*, 9(2), pp. 27–40. Available at: https://doi.org/10.3316/QRJ0902027.

CyberSecPro (2025) 'CyberSecPro'. Available at: https://www.cybersecpro-project.eu/ (Accessed: 9 October 2025).

EDUFI (2025) *Qualifications frameworks | Finnish National Agency for Education*. Available at: https://www.oph.fi/en/education-and-qualifications/qualifications-frameworks (Accessed: 1 December 2025).

ENISA (2022) *European Cybersecurity Skills Framework (ECSF) | ENISA*. Available at: https://www.enisa.europa.eu/topics/skills-and-competences/skills-development/european-cybersecurity-skills-framework-ecsf (Accessed: 1 December 2025).

European Parliament (2017) *C_2017189EN.01001501.xml*. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=oj:JOC_2017_189_R_0003 (Accessed: 1 December 2025).

Ezeah, H.J. (2024) 'Big Data and Privacy with a focus on Statistical Approaches to Ensuring Data Confidentiality', *NEWPORT INTERNATIONAL JOURNAL OF CURRENT ISSUES IN ARTS AND MANAGEMENT*, 4(3), pp. 72–74. Available at: https://doi.org/10.59298/NIJCIAM/2024/4.3.7274.

Forood, A.M.K. *et al.* (2024) 'Advancements in health information technology and their role in enhancing cancer care: Innovations in early detection, treatment, and data privacy', *GSC Advanced Research and Reviews*, 21(1), pp. 228–241. Available at: https://doi.org/10.30574/gscarr.2024.21.1.0380.

Grobler, M., Gaire, R. and Nepal, S. (2021) 'User, Usage and Usability: Redefining Human Centric Cyber Security', *Frontiers in Big Data*, 4. Available at: https://doi.org/10.3389/fdata.2021.583723.

HL7 (2025) *Welcome to the HL7 FHIR Foundation*. Available at: https://www.fhir.org/ (Accessed: 1 December 2025).

Koutras, D. *et al.* (2024) *D3.3 CyberSecPro Portfolio of Cybersecurity Curricula Targeted to Health*. Available at: https://www.cybersecpro-project.eu/wp-content/uploads/2024/06/D3.3_CyberSecPro_Health_vFinal5_12_submitted.pdf (Accessed: 1 December 2025).

ManagiDiTH (2025) 'ManagiDiTH – Master of Managing Digital Transformation in the Health Sector'. Available at: https://managidith.eu/ (Accessed: 25 November 2025).

NIST (2013) 'Cybersecurity Framework', *NIST* [Preprint]. Available at: https://www.nist.gov/cyberframework (Accessed: 1 December 2025).

Priestman, W. *et al.* (2019) 'Phishing in healthcare organisations: threats, mitigation and approaches', *BMJ Health & Care Informatics*, 26(1). Available at: https://doi.org/10.1136/bmjhci-2019-100031.

Rajamäki, J. *et al.* (2024) 'Enhancing Cybersecurity Education for the Healthcare Sector: Fostering Interdisciplinary ManagiDiTH Approach', in *2024 IEEE Global Engineering Education Conference (EDUCON). 2024 IEEE Global Engineering Education Conference (EDUCON)*, pp. 1–7. Available at: https://doi.org/10.1109/EDUCON60312.2024.10578769.

Rajamäki, J., Nepal, A. and Chalkias, I. (2025) 'Implementing CTI Exchange: A Framework for the DYNAMO Project Pilot Phase', *European Conference on Knowledge Management*, 26(1), pp. 709–718. Available at: https://doi.org/10.34190/eckm.26.1.3780.

Rajamäki, J. and Postolache, O.A. (2025) 'Trialogical Learning and Academia-Industry Partnerships in "Sensors for Medical Instrumentation and Signal Processing" Curricular Unit of the ManagiDiTH Master's Program', in M.E. Auer and T. Rüütmann (eds) *Futureproofing Engineering Education for Global Responsibility*. Cham: Springer Nature Switzerland (Lecture Notes in Networks and Systems), pp. 157–168. Available at: https://doi.org/10.1007/978-3-031-85652-5_17.

Rathod, P. *et al.* (2023) 'D2.1 Cybersecurity Practical Skills Gaps in Europe: Market Demand and Analyse'.

Reeves, T. (2006) 'Design research from a technology perspective', in *Educational design research*. Routledge, pp. 64–78. Available at: https://www.taylorfrancis.com/chapters/edit/10.4324/9780203088364-13/design-research-technology-perspective-thomas-reeves (Accessed: 1 December 2025).

Tavares, A., Sousa, P. and Proença, R. (2024) 'Exploring the dynamics between artificial intelligence and cybersecurity in Healthcare', *ARIS2 - Advanced Research on Information Systems Security*, 4(1), pp. 20–34. Available at: https://doi.org/10.56394/aris2.v4i1.44.

Wang, F. and Hannafin, M.J. (2005) 'Design-based research and technology-enhanced learning environments', *Educational Technology Research and Development*, 53(4), pp. 5–23. Available at: https://doi.org/10.1007/BF02504682.

XiA (2025) 'XiA – Xpanding Innovative Alliance'. Available at: https://xia-project.iscte-iul.pt/ (Accessed: 12 November 2025).

Yin, R.K. (2018) *Case study research and applications: design and methods*. Sixth edition. Los Angeles London New Delhi Singapore Washington DC Melbourne: SAGE.

Zhang, J. and Zhang, Z. (2023) 'Ethics and governance of trustworthy medical artificial intelligence', *BMC Medical Informatics and Decision Making*, 23(1), p. 7. Available at: https://doi.org/10.1186/s12911-023-02103-9.

Zhou, Z. *et al.* (2024) 'A statistical approach to secure health care services from DDoS attacks during COVID-19 pandemic', *Neural Computing and Applications*, 36(1), pp. 1–14. Available at: https://doi.org/10.1007/s00521-021-06389-6.

Military Leadership in the Age of Cognitive Warfare

Mikko Salminen
Finish National Defence University, Helsinki, Finland

mikko.k.salminen@mil.fi

Abstract: In many definitions of cognitive warfare, it is highlighted that the target is human mind, and not only the contents of it but the various cognitive processes, such as attention and decision making. The aim in cognitive warfare is often defined to be the influencing of how individuals think, not what they think. Social media is often mentioned as one of the main platforms by which these cognitive influence operations are conducted, possibly enhanced with big data analytics and artificial intelligence (AI) tools. In some definitions it is also included, that in addition to behavioral and cultural sciences, also the exploiting of neuroscience for even more effective and more precisely targeted influences on individual cognition could be considered. The intertwined cognitive, technological, and information dimension form new types of threats which call for a close evaluation of possible effects and countermeasures. The DOTMLPF-I framework (Doctrine, Organization, Training, Materiel, Leadership & Education, Personnel, Facilities, Interoperability) is used in armed forces to guide capability development. This study is a conceptual analysis within these dimensions, with the focus on Leadership and Education dimensions and how they should adapt to face these new types of threats. Military leadership is expected to be fast rational decision-making under uncertainty. However, cognitive warfare challenges this with disinformation and emotional manipulation, for example. Introducing AI functionalities to decision support systems may enhance situational awareness and decision-making speed, but the algorithms may be biased and the systems may operate in a non-transparent way, thus eroding trust towards them. Resilience against effects of cognitive warfare requires training in media literacy and critical thinking, but also in emotional skills. The forming of situational understanding is one of the main targets in cognitive warfare and it should be supported with specialized training, for example by targeting the underlying cognitive processes of forming situational understanding. For this purpose, cognitive training should be targeted to, for example, attention control, working memory, and mental flexibility. In the current manuscript the challenges posed by cognitive warfare to military leadership are reviewed with suggested solutions.

Keywords: Cognitive warfare, Military leadership, Leadership training, Information warfare, Psychological warfare

1. What is Cognitive Warfare?

Cognitive warfare has been defined as a new concept in military sciences and defence studies. Central to various definitions of this concept have been the using of human cognition and new technologies for influencing decision making and behavior (Deppe & Schaal, 2024). The term cognitive refers in the various definitions to influencing human cognitive processes, such as perception, attention, memory, and eventually, decision making. By no means, the concept of influencing adversary thinking is not a new idea and psychological operations have been part of military strategy for ages. The novelty in the concept of the cognitive warfare comes (supposedly) from utilization of recent research on cognitive science and the neurosciences, and from the new technologies. Often artificial intelligence (AI) and social media platforms are mentioned as new technologies, but some scholars have also identified connections to even neurotechnology and nanotechnology (Grigsby et al., 2023; Claverie & Du Cluzel, 2022; Morelle et al., 2023).

Cognitive warfare is presented as some kind of a paradigmatic shift in modern conflicts, targeting the human cognitive processes instead of (merely) influencing opinions. This can be seen in definitions such as the one by Claverie & Du Cluzel (2022), who write that: “The stated objective is to attack, exploit, degrade or even destroy how someone builds their own reality, their mental self-confidence, their trust in processes and the approaches required for the efficient functioning of groups, societies, or even nations.”

The field of study of cognitive warfare is still new and there are varying definitions of the phenomenon. Even if these definitions share some similarities, most of them mentioning cognitive processes and new technologies, there are still differences. This might hinder not only the academic study of the field, but also doctrinal work and preparing. Many scholars and analysts agree on the fact that there are new and developing ways to use information technologies as an attack vector for influencing both decision makers and also the public opinion. These operations are labeled as influence operations, cyber-enabled influence operations, or next-generation psychological warfare (Hoffman, 2025), for example. Advances in AI, big data analytics, and neurotechnology enable various capabilities for attacking and defensive cognitive warfare, ranging from using AI for personalized disinformation campaigns to brain-computer interfaces for monitoring stress responses and thus adapting user interfaces of future systems. Cognitive warfare raises also unprecedented ethical challenges by targeting subconscious processes and potentially violating human autonomy.

It is true that technological developments are constantly offering new and more efficient ways for influencing the decision makers and targeted populations. However, it is debatable whether these developments in technologies are such that they entitle the defining of new type of warfare. It is justified to ask what exactly "cognitive warfare" adds beyond "political warfare" (Armstrong, 2025) or some other concept. In addition, cognitive warfare should be positioned in relation to psychological warfare, information warfare, and psychological operations (Ibrahim et al., 2023). The concept of cognitive warfare should be studied critically, with a focus on forming a sound and useful definition, that is needed for both advancing the academic study of the field but also strategic and operative preparing for these new types of threats (e.g., Bebber, 2025).

An interesting direction for more tangible definition of cognitive warfare may be reached by defining the properties, actions, and operations of cognitive warfare. One such example is the set of three defining features for cognitive warfare actions, proposed by Morelle and colleagues (2023). First, the purpose of the action should be "broader than what is immediately apparent". The authors present an illustrative example where a cyberattack is actually targeted to erode trust towards the bank. Second, the action has a diffuse nature, that is, the action is part of a strategy consisting of various different actions and different means that construct a chain reaction where it is difficult to identify the origin of the attack and to estimate the long-term effects. Third, the target of the actions is cognition. This interesting approach to form a definition offers tangible aspects that can be further identified from actual events from recent history. Even if there is still no standard or a widely accepted definition for cognitive warfare, it is suggested that the different definitions share the formulation that as the methods to influence evolve from traditional propaganda to AI-enhanced influence operations, there is a significant shift in the character of these operations and new approaches and competences may be needed also from military leadership.

This study is a conceptual analysis on how military leadership is challenged by cognitive warfare.

1.1 Ways of Cognitive Warfare

Some recent events that have been labeled in the literature as representing or including elements of cognitive warfare are election manipulation in Baltic countries during 2018-2019, annexation of Crimea in 2014, Russian narrative on repressions of ethnic Russians in Georgia and Ukraine after the annexation of the Crimea, and the Chinese disinformation campaign with the narrative that Taiwan is and will be part of China (Fenstermacher et al., 2023). All these events include actions and operations on various domains, and cognitive warfare is merely a one viewpoint to these totalities. It must also be noted, that the time-line of cognitive warfare doesn't always line with the more traditional or kinetic operations. Means of cognitive warfare may be used to prepare the battlespace, for example there are notions of Russian agents intending to coordinate unrest in Crimea before the annexation (Fenstermacher et al., 2023).

Decision making is a central function in military leadership. In the research literature there have been identified groups of objectives of cognitive warfare that are related to decision-making. These include leading of the enemy into taking a certain course of action, influencing the enemy's decision-making process and thus leading to bad decisions and possibilities to exploit these, weakening the enemy's decisions so that they end up being delayed or lacking initiative, protecting of one's own decision-making from enemy influence, and enhancing one's own capability for decision making (Sattler, 2023). Cognitive warfare methods can be used to attack decision making capability directly or more indirectly by influencing the environment in which the decisions are made, by manipulating troop morale or public opinion, for example (Sattler, 2023). The shaping of the decision space prior to and during the operation has been identified as one of the key factors in modern conflicts (Reczkowski & Lis, 2022). This makes countering cognitive warfare difficult, the indirect effects may be built over longer periods of time and may become visible when it is already too late.

The various methods of cognitive warfare challenge leadership and military decision-making in many ways. For example, the Russian reflexive control, or conveying to the opponent information with a goal to incline them to voluntarily make a pre-determined decision (Thomas, 2004). These influence operations may be difficult to detect, especially if conducted utilizing new technological means, for example by automated social media bot farms to build and spread narratives with memes (e.g., Ferrara et al., 2016; Munk, 2025), possibly using false information and emotionally engineered content for more effective spreading (Brady et al., 2017; Vosoughi et al., 2018;) or using personalized attacks to identified target individuals (Matz et al., 2017), possibly even mobilizing crowds against them (Rahman, 2024). Russian propaganda is typically rapid, continuous and repetitive. This is an effective approach, since first impressions tend to be resilient, and repetition leads easily to familiarity and then to acceptance (Paul & Matthews, 2016).

One way to counter these types of actions is the increasing of training on digital media literacy and critical thinking skills of the staff, and also including to the decision-making process check points for considering alternative courses of action and wargaming their effects.

2. Effects of Cognitive Warfare and how to Mitigate Them

There are already some practices and examples that are proposed to act as defensive mechanisms against of diminish effects of cognitive warfare. These methods can be classified by the DOTMLPF-I framework, which stands for Doctrine, Organization, Training, Materiel, Leadership, Personnel, Facilities, and Interoperability, and is widely used in western militaries when planning to build capabilities (NATO, 2016). Regarding the **Doctrine**, the forthcoming NATO Cognitive Warfare Concept (NATO ACT, 2024) will likely formalize and harmonize the definition of this phenomenon across allied nations and also serve doctrinal development work. In another essay, primary lines of effort were identified for cognitive-security strategic approach (Cheatham et al., 2024). The authors suggested that the U.S. Department of Defence should, through the Joint Staff, lead the work of integrating the cognitive domain to the joint doctrine and also that cognitive warfare training should be included at every echelon.

An example of the **Organization** category is the Swedish Psychological Defence Agency, which was established to coordinate and develop the operations of agencies in psychological defence. Cognitive warfare may affect both the armed forces and also the civilian society, this leads to need for different agencies, departments and ministries to collaborate effectively on the topic and the management structures should support this. One example of building resilience towards wide range of threats, including hybrid threats, is the Finnish comprehensive national defence, a model of cooperation to secure society's vital functions mobilizing the whole society (e.g., Sederholm, Rannikko & Salo, 2025).

For **Training** there are examples of various training programs to enhance, for example, digital media literacy skills that build resilience against cognitive warfare (e.g., Nygren et al., 2021). Developments in the **Materiel** category include, for example projects to build AI capabilities in decision support systems, where safety is considered even if there might be strong desire and hurry for ready systems (Probasco et al., 2025). Another, more futuristic example is the combining of nanotechnology, biotechnology, information technology, and cognitive technology for ultimately constructing augmented human operators (Claverie & Du Cluzel, 2022). These types of augmentations may enhance resilience against effects of cognitive warfare, but also open up new types of attack vectors for targeting cognitive processes.

For **Leadership & Education** there are ongoing programs to build soldiers and leaders cognitive resilience (Andoh, 2025). Leaders should be active in informing their team members on the impacts that new technologies and social media-evoked cognitive biases could have and to facilitate training in topics such as mindfulness and cognitive reappraisal techniques. In addition, the leaders should work to increase digital and media literacy training to increase resilience against hostile influence operations (Cheatham et al., 2024).

In the model of Herlihy (2022) the direct approach to cognitive dominance includes brain education, cognitive science-based learning, and managing information overload. Technological interventions are included also. Education and learning should include topics such as metacognition and thinking skills, the author suggests that these could be fostered by offering education on the functioning of the brain and also training in mindfulness skills. For the cognitive science-based learning the author proposes intelligent tutoring and increasing of the tacit knowledge or durable knowledge in long-term memory of the soldier to support critical thinking skills. Managing information overload is suggested to be supported by a change in the culture of "always on", where real-time availability is expected from the leaders, and also by offering education on effects of multitasking, interruptions, information overload, and time-management skills. These changes would affect various levels and functions of military organizations and for a successful change to occur, a change in the culture would be needed, where the promoting of cognitive capabilities are central to achieve cognitive dominance. And it is the task of leaders in all echelons to support and coordinate the building of such a culture.

The increasing digitalization of the warfare may lead to increasing demands for human decision making. Decision making is required to be made at "post-human speed", thus leading to increased demand for AI-based solutions (Crilly, 2022). In addition, Crilly (2022) suggest that in future post-digital hyper-war leaders would increasingly need competencies such as ingenuity, intellect, media-savviness, and legal and digital competencies. Due to these requirements, the author also suggests that younger generations of leaders should be given a more prominent role in leading future wars, implying that more senior leaders may not be

experts in the new type of warfare. It is true that military leadership relies on rapid, rational decision-making under uncertainty. Cognitive warfare undermines this process by introducing misinformation, deepfakes, and emotional manipulation (Qureshi & Khan, 2024). Research demonstrates that emotional disruption significantly impairs judgment, increasing susceptibility to adversarial influence (Bakir & McStay, 2020; Bakir et al., 2024). Targeted narratives can be used to erode trust within command structures, delaying critical decisions. The psychological toll on leaders and troops further exacerbates operational risks. Practicing of decision-making is simulations with induced stress and cognitive challenges has been suggested as a one way to increase decision-making resilience against cognitive warfare (Sattler, 2023).

Situational understanding refers to the ability to perceive, comprehend, and project the status of the operational environment to support decision-making (Endsley, 1995). In military contexts, this skill is essential for mission success and operational adaptability. The concept of situational understanding builds upon Endsley's (1995) model of situation awareness, which includes perception, comprehension, and projection. Military doctrine emphasizes that situational understanding extends beyond awareness by, for example, integrating critical thinking, collaboration, and predictive analysis (U.S. Army, 2019). Teaching situational understanding requires a combination of cognitive and experiential methods. Classroom instruction can introduce theoretical models and decision-making frameworks. Case-based learning and historical scenario analysis help students recognize patterns and anticipate outcomes. Simulation-based training is one of the most effective methods for developing various managerial skills. Virtual and constructive simulations immerse trainees in complex environments, requiring real-time assessment and decision-making (Salas et al., 2009). Live exercises and wargaming further reinforce these skills by replicating uncertainty and operational stress (U.S. Army, 2019). It is suggested that to build resilience against cognitive warfare, there should be dedicated training for situational awareness and understanding, and also for decision making. The guidelines for situational awareness training and education by Endsley and Jones (2011) could be used when creating and evaluating such programs.

Cognitive training programs, such as those focusing on attention control, working memory, and mental flexibility, enhance the underlying processes that support situational understanding (Blacker et al., 2019). Techniques like red teaming and structured analytic methods encourage critical thinking and challenge assumptions. Advances in artificial intelligence and adaptive learning systems enable personalized training for situational understanding. Intelligent tutoring systems can assess learner performance and adjust scenarios dynamically to target weaknesses (Bell & Kozlowski, 2008). Overall, situational understanding is a multifaceted skill that can be taught and trained through a blend of theoretical instruction, experiential learning, and cognitive enhancement. Military organizations should invest in simulation technologies, adaptive learning platforms, and critical thinking programs to prepare leaders for complex operational environments.

For **Personnel** there are developments for continuous cognitive monitoring for every soldier (Gillis, 2024).

In an integrated approach to cognitive dominance Herlihy (2022) presents both indirect and direct approach for expanded cognitive capability and optimized cognitive performance. Indirect approach includes methods such as talent management, and physiological and pharmacological interventions. Regarding the talent management the author suggests more refined methodology for assessing the individuals' cognitive traits and capabilities during recruiting. In addition to more traditional assessment tools also neuroimaging is suggested. For the physiological interventions the author refers to previous studies that have shown the effect that physical training, nutrition, and sleep have on cognitive functions. Pharmacological interventions could be conducted using various types of stimulants, from caffeine to prescribed medicine to increase cognitive performance and alertness during long missions. Wearable devices are offered as one possible solution in supporting behavioral change, for example for better sleeping habits. Other technological interventions include neurotechnology, such as electro-magnetic stimulation to induce better sleep quality.

For **Facilities** there are synthetic training environments for various types of training (Marino, 2025).

Cheatham and colleagues (2024) suggested to develop defensive mechanisms and tools, for example AI tools that would help in identifying patterns of disinformation and to disrupt hostile attempts to influence the information environment. The integration of artificial intelligence into command and control (C2) systems introduces both opportunities and vulnerabilities. AI-driven decision-support systems promise, for example, enhanced situational awareness but raise concerns about algorithmic bias (e.g., Probasco et al., 2025). NATO's cognitive command initiatives aim to turn traditional command and control structure and systems with AI-functionality to a more predictive, adaptive, and resilient system (Großwald Systems, 2025). Finally, **Interoperability** is strengthened by multinational planning tasks and exercises. There are already, for example,

commercial software that can be used in such large multinational exercises for simulating information environment (Bergh, 2023), for more realistic training against threats of cognitive warfare.

3. Discussion

Cognitive warfare represents a transformative challenge for military leadership, shifting the locus of conflict from physical domains to the human mind. Answering to this threat requires theoretical and doctrinal clarity, technological safeguards, dedicated training and education, and also leadership and decision-making resilience. By prioritizing these areas, military organizations can preserve decision-making integrity and operational effectiveness in the era of cognitive warfare.

Cognitive warfare imposes also profound organizational, doctrinal, and psychological challenges. Applying DOTMLPF-I enables comprehensive mitigation through doctrine adaptation, organizational restructuring, training, technological safeguards, and interoperability. Whole-of-government support and continuous wargaming of possible futures are critical to maintaining military effectiveness in an era where war is fought in minds and on battlefields. Professional military education must also evolve to address cognitive threats. There are ethical dilemmas related to influence operations, particularly regarding civilian populations. Balancing operational effectiveness with international law and ethical standards remains a challenge for leadership. Integrating cognitive security modules into education and training can prepare leaders for these complexities. Cognitive warfare exploits psychological vulnerabilities, eroding trust within units and between allies. Deepfake technologies and AI-generated contents may amplify these effects and create constant uncertainty. Building cognitive resilience requires comprehensive training in media literacy, emotional regulation, and critical thinking. Leadership development programs must incorporate these competencies to build resilient decision-making. Organizational resilience also depends on fostering a culture of adaptability.

The convergence of psychological, technological, and informational dimensions in cognitive warfare demands a holistic response. Military leadership must navigate various new and emerging issues. Collaborative frameworks between allies to build cognitive resilience, and effective using of the new technologies for defensive purposes are needed to mitigate risks. Future research should explore standardized definitions, cross-cultural training programs, and ethical frameworks for cognitive operations. Failure to adapt could result in strategic paralysis and erosion of public trust in defense institutions.

State and also non-state actors use cognitive operations for their strategic objectives. In future battleground cognitive dominance can be achieved, for example, by AI-enabled influence operations, targeting both military and civilian populations. Every military leader is a target of cognitive warfare, but so are their subordinates and also the people of the civilian society, with whom the wars are fought. Cognitive warfare poses various old and also new challenges to military leadership. Situational understanding and evaluating of the future become increasingly difficult under the operations of cognitive warfare. Uncertainty of the situation and suspicion towards information slows down decision making and burdens leaders and decision makers. The uncertainty may be evoked between leaders and their subordinates, between military and the civilian society, between allies from different nations, and towards information from various sources, from allies or from own sensors and intelligence functions.

These all developments have an effect to the military leadership, as leaders are the ones leading and coordinating work for preparedness. Every leader has to be familiar with threats from social and digital media. Depending on the hierarchical level or echelon, the threats vary and so do the responsibilities of the leader, whether they are about watching for mobile phone and social media use of the subordinates or the forming and supervising of higher order communication strategies. A shared task and role for leaders in all echelons is to act as mediators of situational understanding and also as pedagogical leaders in coaching the subordinates and team members in their task of becoming more proficient in forming situational understanding to enable efficient decision-making and action.

Ethics and AI declaration: Ethical clearance was not required for the current research according to the national laws and regulations. No artificial intelligence (AI) tools were used for writing this paper.

References

Andoh, E. (2025) Building mental strength in the military. American Psychological Association. Available at: https://www.apa.org/monitor/2025/11-12/mental-strength-military. (Accessed: 26 November 2025).

Armstrong, M. (2025) "Cognitive Warfare" fails the cognitive test. Available at: https://mountainrunner.substack.com/p/cognitive-warfare-fails-the-cognitive. (Accessed: 26 November 2025).

Bakir, V., Laffer, A., McStay, A., Miranda, D., and Urquhart, L. (2024) On manipulation by emotional AI: UK adults' views and governance implications. Frontiers in Sociology, Vol 9, 1339834. https://doi.org/10.3389/fsoc.2024.1339834

Bakir, V., and McStay, A. (2020) Empathic media, emotional AI, and the optimization of disinformation. In Affective politics of digital media (pp. 263-279). Routledge.

Bebber, J. (2025) China is waging cognitive warfare. Fighting back starts by defining it. Defense One. Available at: https://www.defenseone.com/ideas/2025/03/china-waging-cognitive-warfare-fighting-back-starts-defining-it/403886/. (Accessed: 26 November 2025).

Bell, B. S., and Kozlowski, S. W. J. (2008) Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. Journal of Applied Psychology, Vol 93, pp 296-316.

Bergh, A. (2023). Chapter 11 – SOMULATOR: Developing cogwar resilience through social media training. Mitigating and Responding to Cognitive Warfare. NATO STO Technical Report RDP STO-TR-HFM-ET-356.

Blacker, K. J., Hamilton, J., Roush, G., Pettijohn, K. A., and Biggs, A. T. (2019) Cognitive training for military application: a review of the literature and practical guide. Journal of Cognitive Enhancement, Vol 3, No. 1, pp 30-51. https://doi.org/10.1007/s41465-018-0076-1

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., and Van Bavel, J. J. (2017) Emotion shapes the diffusion of moralized content in social networks. Proceedings of the National Academy of Sciences, Vol 114, No. 28, pp 7313-7318. https://www.pnas.org/doi/pdf/10.1073/pnas.1618923114

Cheatham, M. J., Geyer, A. M., Nohle, P. A., and Vazquez, J. E. (2024) Cognitive warfare: the fight for gray matter in the digital gray zone. Joint Force Quarterly, Vol 114, No. 2, pp 83-91. https://digitalcommons.ndu.edu/joint-force-quarterly/vol114/iss2/15

Crilly, M. (2022) Prosecuting the post-digital hyper-war: Preparing for the upcoming war of decision superiority and cognitive dominance. The RUSI Journal, Vol 167, No. 4-5, pp 78-82. https://doi.org/10.1080/03071847.2022.2144431

Claverie, B. and du Cluzel, F. (2022) Cognitive warfare: The advent of the concept of cognitics in the field of warfare. In: B. Claverie et al., eds. Cognitive warfare: The future of cognitive dominance. Brussels: NATO Science and Technology Organization, pp 2-1–2-8. Available at: https://hal.science/hal-03635889/document. (Accessed: 26 November 2025).

Deppe, C., and Schaal, G. S. (2024) Cognitive warfare: a conceptual analysis of the NATO ACT cognitive warfare exploratory concept. Frontiers in Big Data, Vol 7, 1452129. https://doi.org/10.3389/fdata.2024.1452129

Endsley, M. R. (1995) Toward a theory of situation awareness in dynamic systems. Human Factors, Vol 37, No. 1, pp 32–64.

Endsley, M. R., and Jones, D. G. (2025) Designing for situation awareness: An approach to user-centered design. CRC press.

Fenstermacher, L., Uzcha, D., Larson, K., Vitiello, C., and Shellman, S. (2023) New perspectives on cognitive warfare. In Signal Processing, Sensor/Information Fusion, and Target Recognition XXXII, Vol. 12547, pp 172-187. https://doi.org/10.1117/12.2666777

Ferrara, E., Varol, O., Davis, C., Menczer, F., and Flammini, A. (2016) The rise of social bots. Communications of the ACM, Vol 59, No. 7, pp 96-104. https://dl.acm.org/doi/pdf/10.1145/2818717

Gillis, W. (2024) Fort Moore launches Cognitive Monitoring Program to enhance Soldier health and readiness. U.S. Army. Available at: https://www.army.mil/article/279247/fort_moore_launches_cognitive_monitoring_program_to_enhance_soldier_health_and_readiness. (Accessed: 26 November 2025).

Grigsby, C. C., McKinley, R. A., Bridges, N. R., and Carpena-Núñez, J. (2023) Developing cognitive neuroscience technologies for defense against cognitive warfare. Mitigating and Responding to Cognitive Warfare. NATO STO Technical Report RDP STO-TR-HFM-ET-356.

Großwald Systems. (2025) Cognitive command: AI-enhanced decision-making in Multinational Battle Networks. Available at: https://www.grosswald.org/cognitive-command-ai-enhanced-decision-making-in-multinational-battle-networks/. (Accessed: 26 November 2025).

Herlihy, D. J. (2022) Cognitive performance enhancement for multi-domain operations. The US Army War College Quarterly: Parameters, Vol 52, No. 4. https://doi.org/10.55540/0031-1723.3188

Hoffman, M. (2025) Assessing "Cognitive Warfare". Small Wars Journal. Available at: https://smallwarsjournal.com/2025/11/14/assessing-cognitive-warfare/. (Accessed: 26 November 2025).

Ibrahim, F., Rhode, S., and Daseking, M. (2023) A systematic review of cognitive and psychological warfare. Available at: https://tdhj.org/blog/post/cognitive-psychological-warfare/ (Accessed: 26 November 2025).

Marino, C. (2025) Reality check. U.S. Army. Available at: https://www.army.mil/article/286728/reality_check. (Accessed: 26 November 2025).

Matz, S. C., Kosinski, M., Nave, G., and Stillwell, D. J. (2017) Psychological targeting as an effective approach to digital mass persuasion. Proceedings of the National Academy of Sciences, Vol 114, No 48, pp 12714-12719. www.pnas.org/cgi/doi/10.1073/pnas.1710966114

Morelle, M., Julien, C., Marion, D., and Jean-Marc, A. (2023) Towards a definition of cognitive warfare. In Conference on Artificial Intelligence for Defense. Available at: https://hal.science/hal-04328461/document. (Accessed: 26 November 2025).

Munk, T. (2025) Digital Defiance. Memetic Warfare and Civic Resistance. European Journal on Criminal Policy and Research, Vol 31, pp 501-528. https://doi.org/10.1007/s10610-025-09613-4

NATO Allied Command Transform. (2024) Allied Command Transformation develops the Cognitive Warfare Concept to combat disinformation. NATO Allied Command Transformation. Available at: https://www.act.nato.int/article/cogwar-concept/. (Accessed: 26 November 2025).

NATO, Joint Analysis and Lessons Learned Centre (2016). Joint Analysis Handbook. Available at: https://www.jallc.nato.int/application/files/9416/0261/6056/Joint_Analysis_Handbook_4th_edition.pdf. (Accessed: 26 November 2025).

Nygren, T., Guath, M., Axelsson, C.-A. W., and Frau-Meigs, D. (2021) Combatting visual fake news with a professional fact-checking tool in education in France, Romania, Spain and Sweden. Information, Vol 12, No. 5. https://doi.org/10.3390/info12050201

Paul, C., and Matthews, M. (2016) The Russian "Firehose of Falsehood" Propaganda Model. RAND Corporation. Available at: https://www.rand.org/pubs/perspectives/PE198.html (Accessed: 26 November 2025).

Probasco, E. S., Toner, H., Burtell, M., and Rudner, T. G. (2025) AI for Military decision-making. Center for Security and Emerging Technology. Available at: https://cset.georgetown.edu/wp-content/uploads/CSET-AI-for-Military-Decision-Making.pdf. (Accessed: 26 November 2025).

Qureshi, J., and Khan, S. (2024) Deciphering deception: The impact of AI deepfakes on human cognition and emotion. Journal of Applied Artificial Intelligence, Vol 2, No. 1, pp 101-107.

Rahman, E. A. (2024) Indirect Swarming and Its Threat to Democracies: A New Frontier in Online Harassment. New America. Available at: https://www.newamerica.org/oti/briefs/indirect-swarming-and-its-threat-to-democracies/. (Accessed: 26 November 2025).

Reczkowski, R., & Lis, A. (2022) Cognitive Warfare: what is our actual knowledge and how to build state resilience?. Bezpieczeństwo. Teoria i Praktyka, 48(3), 51-61. https://doi.org/10.48269/2451-0718-btip-2022-3-003

Salas, E., Wildman, J. L., and Piccolo, R. F. (2009) Using simulation-based training to enhance management education. Academy of Management Learning & Education, Vol 8, No. 4, pp 559-573.

Sattler, T. (2023) The impact of cognitive warfare on strategic decision making in NATO. In Pashova et al. (ed.), CMDR COE Proceedings, pp 177-205.

Sederholm, T., Rannikko, R., & Salo, M. (2025) Total defence model at the heart of Finland's national defence and resilience. In Rongved (ed.), European Total Defence (pp. 115-134). Routledge. https://doi.org/10.4324/9781003497370

Thomas, T. (2004) Russia's reflexive control theory and the military. Journal of Slavic Military Studies, Vol 17, No. 2, pp 237-256. https://doi.org/10.1080/13518040490450529

U.S. Army. (2019). Army Doctrine Publication 6-0: Mission Command. Department of the Army.

Vosoughi, S., Roy, D., and Aral, S. (2018) The spread of true and false news online. Science, 359(6380), 1146-1151. https://doi.org/10.1126/science.aap9559

Evaluation of AI Agent Accelerated Cyber Operations Planning

Pietari Sarjakivi and Panu Moilanen
University of Jyväskylä, Finland

pietari@sarjakivi.fi
panu.moilanen@jyu.fi

Abstract: As societies become increasingly digital, they are more exposed to cyber threats that have the potential to harm human life and damage critical infrastructure and other assets. To counter these fast-paced threats, Defensive Cyber Operations (DCO) leaders must enhance their capabilities for rapid decision-making and response. Artificial Intelligence (AI), as a radical levelling technology, has the potential to accelerate DCOs; however, the existing solutions frequently focus on narrow technical use-cases and lack emphasis on the leadership dimension of DCO. The purpose of this paper is to address that gap by researching how AI can accelerate one of the most relevant DCO use-cases identified in the author's earlier research, course of action recommendation, especially in operations planning. The study is based on case study methodology, where AI agent-generated operations plans are compared to real-life DCO operations plans made by leading experts in the world's most complex defensive cyber exercise, NATO Locked Shields 2025. The study focuses on two of the most critical decisions a DCO leader needs to make during the primary process of DCO: Prioritization of defended capabilities and assets, and the right resourcing and allocations. The selected exercise provided an excellent platform for this study to compare multiple human-made plans to machine-made plans, as 17 world-class blue teams were given the same exercise scenario and operation order. As a result, this paper demonstrates that with proper architecture and context engineering, AI can significantly accelerate DCO leaders' decision-making in operations planning, while human-machine teaming is still needed to navigate a complex operating environment where cyber operations are typically conducted. The main contributions of this paper are 1. evaluation of an AI agent's performance in DCO operations planning in comparison to human experts, and 2. construction of a reference architecture for the DCO planning agent. Future research can be built to improve the results of the reference architecture. As AI's capabilities are developing rapidly, it is expected that the capabilities of autonomic AI agents will increase.

Keywords: Defensive cyber operations, Artificial intelligence, Operations planning, AI agent, Context engineering

1. Introduction

With the world digitalizing, societies are becoming more dependent on their digital infrastructure. Simultaneously, cybercrime increases year-by-year and targets societies essential services, like critical manufacturing and healthcare (Federal Bureau of Investigation, 2025). Cyber Operations (CO) are becoming an increasingly integral part of hybrid warfare, with governments' ability to deny their participation on cyberattacks. Critical infrastructure providers, who are often private companies (Azrilyant et al., 2022), have difficulties defending their assets against advanced attackers, and therefore, governments need to provide assistance in the form of Defensive Cyber Operations (DCO) to recover from a cyber incident, like in Minnesota 2025 (City of Saint Paul, 2025). To enable effective DCOs performed by both military and civilians, a joint operations model is essential.

Artificial Intelligence (AI) is a radical leveling technology: it makes cyberwarfare fighting capabilities accessible to every nation and terrorist organization. Therefore, it has the potential to disrupt linear development in multiple domains, including cybersecurity (Snow, 2015). AI can improve COs in at least 22 use cases (Sarjakivi, 2025), and the rapid development of Large Language Models (LLMs) has enhanced AI agent maturity to a level where agents have the capability to autonomously exploit zero-day vulnerabilities in real-life systems with just a few-dollar cost (Fang et al., 2024). AI agents extend LLMs' reasoning and interpretation skills by synergizing Reasoning and Acting (ReAct) skills (Yao et al., 2023). Context engineering, a systematic approach exceeding the limitations of prompt engineering, is used for improving LLM performance (Pajo, 2025).

This study focuses on evaluating AI agents' capabilities to accelerate DCOs. The paper is organized as follows: Section 2 introduces the research methodology, section 3 presents the results for research questions, and section 4 concludes the study with future research topics.

2. Methodology

2.1 Research Methodology

This study utilizes a case study approach and consists of comparative content analysis and structured expert interview phases. A case study is an especially suitable methodology for this research as it focuses on "how" questions in contemporary phenomena within a real-life context where the researcher has little control over the events (Yin, 2003). *Course of Action (COA) recommendation* is the third most important AI use cases

accelerating COs, and, as the other high ranking use cases, namely *information sharing and reporting*, and *data fusion, enrichment, and visualization*, are relatively generic, this CO focused use case was selected for the evaluation. According to the same study, the most important decisions a DCO leader needs to make within the primary process are *prioritization of defended capabilities and assets*, and *the right resourcing and allocations* (Sarjakivi, 2025).

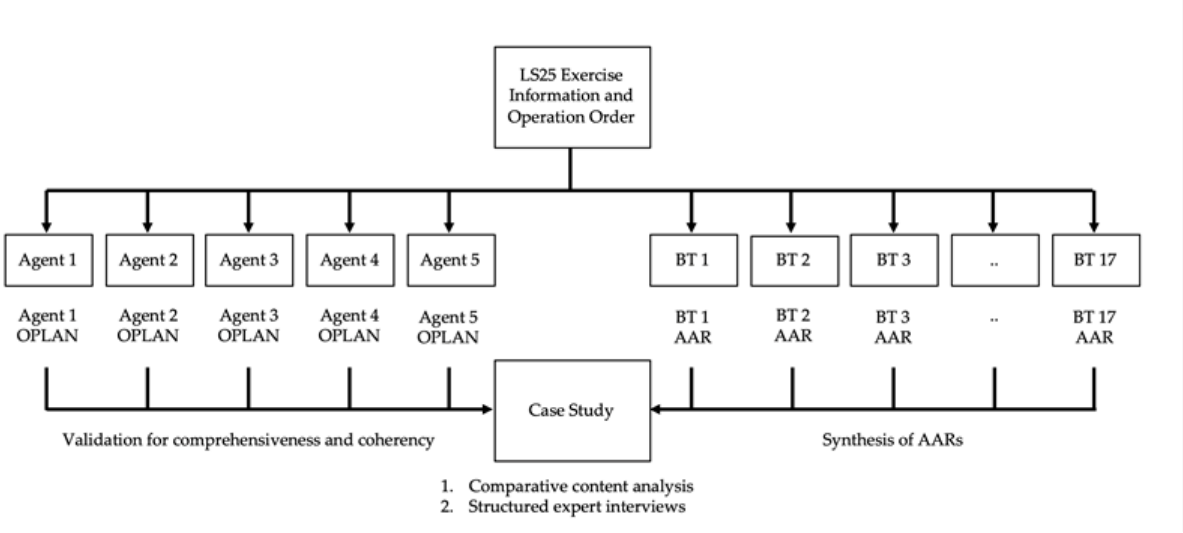


Figure 1: Research methodology utilized in this case study

As shown in Figure 1, the study was conducted by comparing AI agent-generated operations plans to real-life DCO operations plans made by human experts in NATO Locked Shields 2025. Locked Shields is the world's most complex cyber defense exercise, where over 4000 people from 41 nations defend a fictional country's military systems and critical infrastructure from cyber-attacks (NATO Cooperative Cyber Defence Centre of Excellence, 2025). In the exercise, each of the 17 world-class Blue Teams (BT) were given the same exercise information with scenario, rules of engagement, and the Operations Order (OPORD) to defend the environment. Although the BT's Operations Plans (OPLAN) were not available for comparison, most of the teams explain their approach in the After Action Report (AAR), from which the most critical decision can be extracted. Together with its scale and accessibility of AAR information, this exercise provided an excellent platform for comparison of human- and machine-made OPLANs.

To understand the impact of architecture and context engineering, 5 increasingly capable agents were created.

- Agent 1: Simple prompting.
- Agent 2: Advanced prompting using best practices
- Agent 3: Advanced prompting and Retrieval-Augmented Generation (RAG) datastore.
- Agent 4: Advanced prompting, RAG datastore, and manually injected context into the system prompt.
- Agent 5: Advanced prompting, RAG datastore, manually injected context into system prompt, and response improved with the recommendation of a separate Wargamer tool.

Each agent was tested with 3 LLM models, each possessing unique advantages and limitations, such as the needed computing power and the possibility of operating in offline environments. OpenAI's latest and highest performing model, GPT-5, was selected based on its reference test results (OpenAI, 2025b) to represent high-end, centrally hosted closed source models. Open AI's GPT-OSS-120b, which performs in similar level of closed source models like o4-mini (OpenAI, 2025a), was selected as heavy weight open source model that can be run offline as long as dedicated hardware is available, and DeepSeek's latest open source reasoning model, DeepSeek-R1-0528, as a high performing alternative to OpenAI's models (DeepSeek, 2025). As AI agents output different responses with each execution, each agent-model combination was executed 3 times, and the most comprehensive and coherent response was passed to further analysis. In total, 15 responses were analyzed. In the test setup, all the data and processing were done in a purpose-built, high-security environment to ensure secure data handling requirements.

In the first phase, comparative content analysis was performed by manually extracting key elements from AI agent responses and comparing them against the synthesized BT AARs results. The second phase was structured expert interviews with 3 experienced professionals who have experience of leading their nations'

Locked Shields BTs and creating the OPLAN for their teams. Experts evaluated agent responses by scoring the AI-generated OPLAN on a scale of 1-10 and answering the following questions with a scale of Yes (1) – Partially (0,5) – No (0).

- Is this plan executable?
- Is the prioritization of defended capabilities and assets reasonable?
- Are the resourcing and allocations reasonable?

2.2 Research Questions

This study aims to answer the following questions:

- RQ1: Can AI agents improve DCO operations planning?
- RQ2: What kind of context engineering and architecture are required for a DCO AI agent?

3. Results of Research Questions

3.1 RQ1: Can AI Agents Improve DCO Operations Planning?

As shown in Figure 2 below, the overall results of expert analysis indicate that AI agents improve DCO operations planning. Agent 4, which performed the highest, OPLAN scored 7/10 and got 80% from the executability measurement. Although the results are not at a sufficient level for autonomous operations, an AI agent can significantly accelerate operations planning. According to the results, the performance of different models and agents varies heavily, while GPT-5 and gpt-oss-120b seem to perform the highest. Also, the inconsistency between executions remained high throughout the evaluation, highlighting the need of Human-Machine Teaming (HMT), where human operators' analysis of the agents' responses and steer agent response generation to match the needs of a complex operating environment where COs are typically conducted.

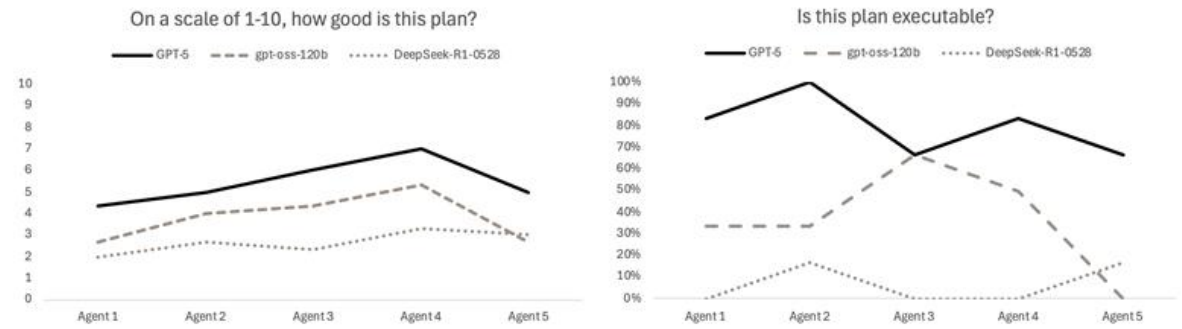


Figure 2: Expert analysis results on the overall quality of AI-generated OPLAN

On specific decisions of *prioritization of defended capabilities and assets*, AI agents seem to perform relatively good, as shown in Figure 3. Agent 4 with the GPT-5 model got a 100% rating from experts, while others performed relatively well. It is noticeable that all other agents hallucinated some of the defended capabilities and or provided incomplete lists. When agent 4's response was given to agent 5 to wargame and improve, the agent 5 focused on partially hallucinated technical details and lost the complete listing of defended capabilities. This is assumed to be caused by overwhelming context and loss of focus on the original task.

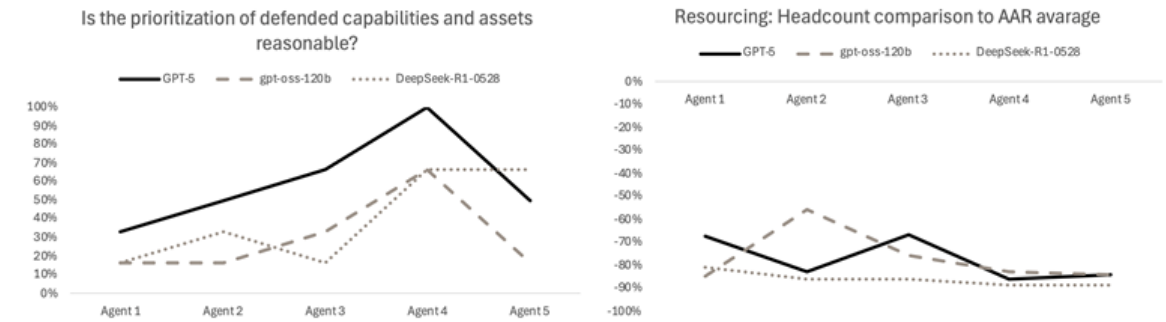


Figure 3: Agents' performance on the most important decisions

The act of choosing *the right resourcing and allocations* was proven to be a difficult task for the agents. As shown in Figure 3, agents choose team size on average 85% below the average of Locked Shields 2025. This variation can be partially explained by the fact that LLMs' training data is based on historical data, while the number of team members in Locked Shiels has over ten folded during the last ten years (NATO Cooperative Cyber Defence Centre of Excellence, 2025; Smeets, 2022). Also, the human-made plan might be impacted by external factors, like the economy and the availability of specialists. According to the AAR synthesis, a larger number of specialists does not guarantee success in the exercise, and in fact, high-performing BTs' participant count was 23% below the average.

Only 33% of AI agent-created organizations were utilizing matrix or defended capability-based team structures, which was the most utilized approach in Locked Shields 2025. Over 93% of agent-generated organizations were built to be low hierarchy, which is contradicting to real-life BT structures, where over 81% utilized a multi-layer organization structure with operations leaders between blue team commanders and squad leaders leading specialists. One contributing element to this variation can be a substantial difference in headcount, as bigger organizations usually need a higher level of hierarchy.

To conclude the study in a reasonable timeframe and make the results comparable, certain limitations had to be made. Agents' technical implementation and prompt tuning were the same for all the models and, therefore, not optimized for any of them. BT's Standard Operating Procedures (SOPs) were not given to the agent, as they vary between BT. Data captured from AARs didn't provide a complete understanding of BTs' OPLANs, and even if they had provided, a single right answer does not exist. While the number of experts analysing AI agents' responses was limited, their analysis was relatively cohesive. Even with limitations identified above, the authors believe the evaluation gives a trustworthy answer to AI agents' DCO operations planning capabilities.

3.2 RQ2: What Kind of Context Engineering and Architecture are Required for a DCO AI Agent?

This construction is built by applying AI agent best practices on Sarjakivi et al's (2025) DCO decision-making model. Reference implementations like military COA creation (Goecks & Waytowich, 2024), disaster response (Goecks & Waytowich, 2023), and information operations (Kereopa-Yorke, 2024) have been reviewed, and lessons learned from those implementations are adopted into this implementation.

3.2.1 Human-machine teaming and right expectations

Although in the academic literature, AI's role in accelerating COs is often seen as an independent agent (39%) (Sarjakivi, 2025), HMT, with its objective to find optimal performance of the team, builds on collaboration and interdependent teamwork of human and machine operators (e.g., AI agents) (Johnson & Bradshaw, 2021). Festor et al (2021) defined 5 levels of autonomy in safety-critical decision-making, where machine operators' role incrementally increases from advisor to independent decision maker while human operators' role decreases from decision maker to assurance and backup role, and eventually drops out of the decision-making loop completely. According to the situational leadership model, human operators' maturity in making decisions without assistance from more experienced practitioners is not universal but dependent on the situation and decision at hand (Hersey et al., 2013). The same model should be applied to machine operators' decision-making. Similarly to delegating decision-making to a junior colleague, the key is to understand how much errors and delays the particular decision tolerates, what the cost of a wrong decision is, and can we implement supporting controls or mitigations in case of a wrong decision made. Additionally, the situation where a decision is made impacts on the autonomy that should be granted to machine operators, and therefore, autonomy should be seen more as a dynamic slider controlled by an experienced human operator.

In the early levels of autonomy, machine operators pre-process and analyse the data needed for decision-making, increasing the speed of the decision-making loop. The scale and speed provided by the machine operator enable more thorough analysis with an extended number of alternative COAs analysed within the same time window provided by the frame operation. With mature enough machine operators, the current dependency on experienced human operators decreases, providing organizations the possibility to conduct multiple overlapping operations simultaneously with fewer or less experienced human operators. According to the results of RQ1 in section 3.1, human cognition is needed to verify the results of machine created plans. Human operators should focus on problems caused by physical and human elements in the plan execution, like extensive delays in responses, inaccuracy in communications, and lack of visibility the events in the real life.

3.2.2 Context engineering and safeguarding best practices

Improving the results of AI agents' performance requires several considerations.

Model. Each model has their unique performance characteristics, and it is critical to choose the right model that fits the purpose (Mbaiossoum et al., 2025). Multimodal models support different data types such as audio and images, base models are suitable for tasks requiring generic knowledge, and fine-tuned models excel on specific tasks. Chat models are optimal for conversation and writing tasks, while reasoning models are better at problem-solving and decision-making. An agent can be built to utilize different models based on the task at hand, for example, using a lightweight model for simple and pre-processing related tasks and premium model for verification and quality assurance. Execution speed and resource or monetary cost of running the model are important considerations for agents. Each model has different knowledge cut-of-date that might impact thinks like understanding of the latest technologies. Openness and hosting arrangements of the model have an impact on privacy, and pre-existing safeguards set limitations on what purpose the model can be used for. Trust towards the model creator and political reasons might contribute to model selection as well. While the open source model available for hosting locally are getting closer or even exceeding the performance of proprietary models in certain tests like high school math, the proprietary models are still outperforming open source models in most of the tests (OpenAI, 2025a; Vellum, 2025).

Prompting. Right prompting is often seen as the primary method for improving agents' performance, and numerous prompting best practices and guides have been published for different purposes (Černý, 2024; Cowan et al., 2023; Ekin, 2023; MacCallum & Lee, 2025; J. Wei et al., 2023). Although a good prompt is specific for the purpose, a synthesis of the mentioned guides provides a description of characteristics for a good prompt.

- A clear task, role, and tone instruct LLMs on what the human operator wants from it.
- The right temperature and guidance for hallucination. LLM needs to be told what to do when it doesn't know the answer, or make it ask for more instructions or data instead of guessing the answer.
- Examples and few-shot prompts steer LLM to understand what is needed from it and enable in-context learning.
- Structured output improves the possibility of chaining and further develops the result. For example, asking LLM to provide an answer in a table with specific column headers ensures that LLM completes all the fields.
- Delimiting and clearly distinct parts of the prompt from each other, making the instruction clearer and preventing certain types of prompt injections.
- Specified step-by-step instructions and Chain-of-Thought prompting where LLM is asked to "think" before it gives the final answer. Ask LLM to provide the reason why it ended up in that response.
- Persistence within agentic conversations. It is important to make the LLM understand that it should wait for other agents until all the conditions are satisfied before returning to the human operator.

Context and tools. Context engineering increases focus and speed of the agent by writing, selecting, compressing, and isolating just the right context for the agent (LangChain, 2025). In addition to instructing a specific focus on a prompt, external information can be provided for LLM by using the RAG technique. Using methods like Model Context Protocol (MCP), agents can invoke tool calls, bringing them capabilities like a calculator, clock, databases, and web search, and enabling them to communicate with other agents (Gan & Sun, 2025). Context window optimization, one of the key component of context engineering (Pajo, 2025), becomes relevant when conversation history and external data exceed build-in limits of the model, resource consumption restrictions, or expected response times. As the RQ1 results demonstrate, manually injecting context simply by attaching only the relevant data into the system or user prompt increases the performance of the agent while keeping the context window optimized.

Architecture. Simplicity and linearity improve long-running agents' performance and minimize out-of-sync related context problems, where agents don't know what other agents in the network are doing (Yan, 2025). Parallelism is needed when agents need highly specialized roles or are initiated based on various events in the operating environment. A multi-agent approach enables agents to have specific instructions and use an optimal LLM model for the task, which improves the collective performance of the agent network. Sharing of context and situational awareness can be done using centralized components. Also, parallel agent operations have their use cases, for example, creating multiple COAs and then rank each other's responses.

Safeguards. AI adversarial attacks and control bypassing tactics have been studied extensively (Greshake et al., 2023; Qi et al., 2023; Wan et al., 2023; A. Wei et al., 2023; Zou et al., 2023) and legislations like EU AI ACT mandate adversarial testing performed for high-risk AI systems (European Union, 2024). While tools like PyRIT (Microsoft AI Red Team, 2024) and SPLX (SPLX, 2025) provide a comprehensive set of automated tests, at least the following safeguards should be implemented.

- Input query moderation, as end users or other agents can control the user prompt.
- Implementing delimiters into the system prompt provides safeguards against prompt injections.
- Output response moderation prevents outbound lateral movement and data leaks.

Comprehensive implementation of the safeguards above is difficult. Agents' access to secret data, its exposure to untrusted content, and its ability to communicate externally form a lethal trifecta, which should be mitigated by eliminating one of the three elements (Willison, 2025).

3.2.3 Agent architecture construction

Essential components of a DCO AI agent are introduced in Figure 4. Human operator communicates with AI agent network using the user interface, storing and accessing documents from the datastore, reviewing and updating the status of tasks in the task list, and reviewing the operation status from event logs. Agent infrastructure consists of an agent framework orchestrating the whole agent network, reasoning and chat models used by agents to comprehend tasks, embedding models for converting documents in the datastore into vector representation, which can be understood by reasoning and chat models, and a vector database for storing and searching vectors. Several tools are used to offload task specific instructions from agents and to standardize activities like unified format event logging and embedding multi-format data into a vector database. Regardless of multiple components in the architecture, all the agent infrastructure can be hosted in a single computer, enabling operations in an isolated environment, for example, in military operations.

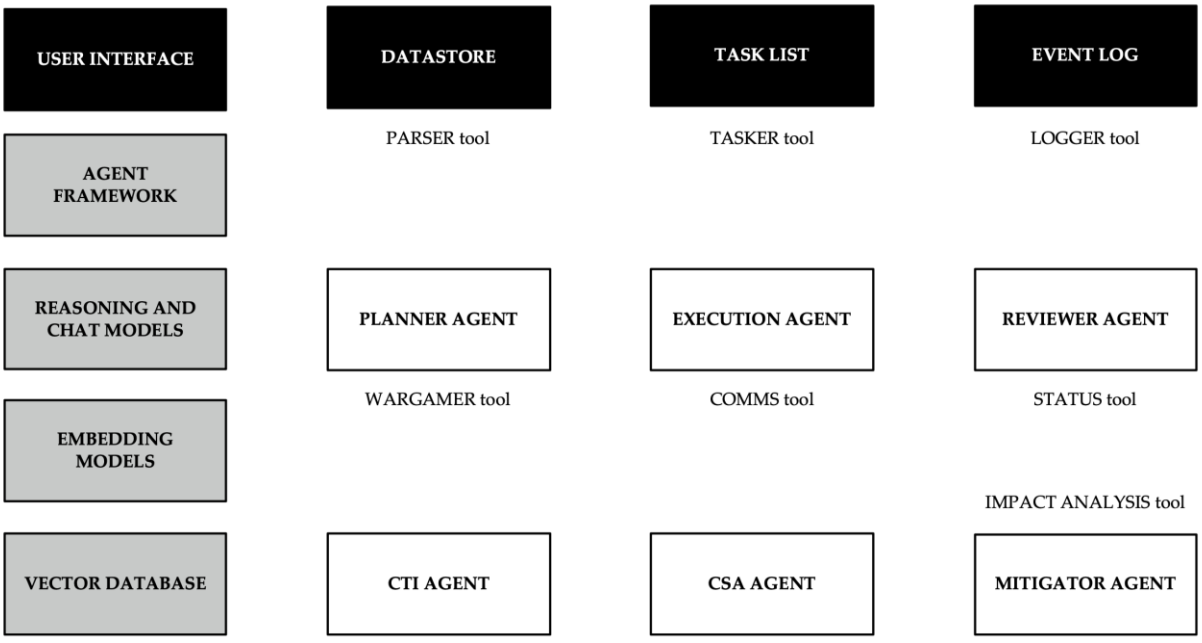


Figure 4: Essential components of DCO AI agent

The construction of the DCO AI agent architecture is introduced in Figure 5. In this architecture, each of the 6 agents has their specific role, while they share tools and resources, and are capable to communicate with each other.

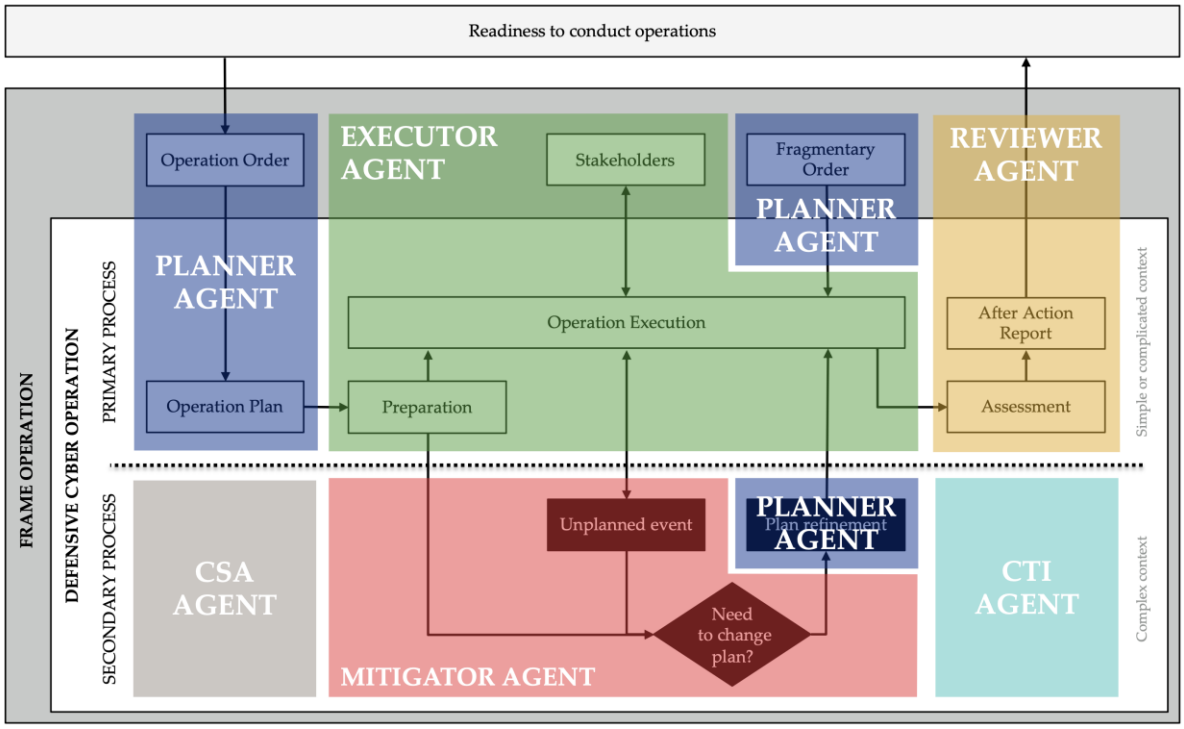


Figure 5: DCO AI architecture construction built on Sarjakivi et al (2025) DCO decision-making process

The *planner agent* is responsible of creating and updating the Operations Plan (OPLAN) based on higher headquarter issued Warning Order (WARNO), Operations Order (OPORD), or Fragmentary Order (FRAGO). Those orders are stored in the *datastore* and should include the objectives, area of operations, timeline, relation, and dependencies to frame operation, Rules of Engagement (ROE), and available support (Wade, 2023). As an agent, using a foundation model as its core, which is not pre-trained for a particular task, needs to understand the context where the operation is performed to create an actionable OPLAN fit for purpose. The *planner agent* uses general operations planning principles, like Military Decision-Making Process (MDMP) (Wade, 2023) and NATO cyberspace doctrine (NATO, 2020) as a base, but scenario and threat landscape, SOPs, specific doctrines, capabilities of the team, and lessons learned from previous operations needs to be context engineered into the agent. In addition to information included into prompts and found from the *vector database* and *event log*, the *planner agent* can call the *Cyber Threat Intelligence (CTI) agent* to understand the threat landscape and the *Cyber Situational Awareness (CSA)* agent to understand the current operational status impacting planning. *Planner agent* uses the *Wargamer tool* to challenge and improve its plans. *Wargamer tool* benchmarks the plan to alternative plans and plans from past operations and tries to find weaknesses from the plan by projecting the enemy responses to activities planned.

The *execution agent* is responsible for ensuring that the team enters the operations well-prepared and that the operation is executed smoothly. It divides OPLAN into tasks in the *task list* using the *tasker tool* and the *communications tool* to communicate with external stakeholders. Using the *status tool*, agent follows operation progress by comparing OPLAN to *event log* and task completion status and calls *CTI* and *CSA agents* for more information if needed. If part of the operation enters the secondary process in the DCO decision-making model, the *executor agent* calls the *mitigator agent* to bring the operation back to the primary process. Both *executor* and *mitigator agents* use the *impact analysis tool* to understand the impact of a cyber event on the overall operation. At the end of the operations, the *reviewer agent* assesses the operation based on the *event log* and the *status tool* and creates an AAR. AAR, together with identified lessons learned, are stored into the *datastore* and used to update SOPs, thus improving future operations performance.

4. Discussion

As a response to the national critical infrastructure's growing exposure to fast-evolving cyber threats, cyber defenders must improve their capabilities. This study demonstrated that AI agents can significantly accelerate decision-making in the most critical decisions DCO leaders need to make, namely, *prioritization of defended capabilities and assets*, and *the right resourcing and allocations.* Prioritization and generic plan generation scored high in the evaluation, and agents could generate out-of-the-box ideas and COAs quickly, supporting

human operators' innovation. While proper context engineering improves the quality of the agent's results, the HMT is needed as agents tend to hallucinate and provide varying results, for example, to fill all the cells in a table. Giving more details to the agent makes it focus on other details, leading to reduced focus on the original objective. Human oversight is needed, especially on resourcing-related tasks, as the agent didn't perform particularly well in that domain.

Even with rapid development with LLMs, AI agents are not ready for autonomously execute DCOs. Well-functioning HMT is needed to navigate a complex operating environment where COs are typically conducted. The asynchronous nature of cyber physical environments where humans and machines operate seem to be difficult for LLMs to comprehend.

This study provided a construction of a reference architecture for the DCO planning agent. Future research can be built to improve the results of the reference architecture. As AI's capabilities are developing rapidly, it is expected that the capabilities of autonomic AI agents will increase in the near future.

Acknowledgements

The authors would like to thank the Finnish Ministry of Defence for its support in writing this article (SOW275).

AI declaration: While producing this article, AI was used to improve the spelling and to perform a grammar check. As the studied target was an AI agent, the evaluated product naturally includes AI components.

Ethics declaration: No ethical clearance was needed for this study.

References

Azrilyant, J., Sidun, M., & Dolashvili, M. (2022) Fact and Fiction: Demystifying the Myth of the 85%, *The George Washington University Elliott School of International Affairs*.

Černý, J. (2024) Prompt Engineering: Tactics and Techniques in Open-Source Intelligence, *Journal of Information Warfare,* 23(3).

City of Saint Paul (2025) "Digital Security Incident Info Hub", [online], https://www.stpaul.gov/departments/emergency-management/digital-security-incident-info-hub.

Cowan, L. S., Lerman, D. C., Berdeaux, K. L., Prell, A. H., & Chen, N. (2023) A Decision-Making Tool for Evaluating and Selecting Prompting Strategies, *Behavior Analysis in Practice*, 16(2), 459–474.

DeepSeek. (2025) "DeepSeek-R1-0528 Release", [online], https://api-docs.deepseek.com/news/news250528.

Ekin, S. (2023) Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices. *Institute of Electrical and Electronics Engineers (IEEE)*.

European Union. (2024) Regulation (EU) 2024/1689 of the European Parliament and of the Council.

Fang, R., Bindu, R., Gupta, A., & Kang, D. (2024) LLM Agents can Autonomously Exploit One-day Vulnerabilities, arXiv preprint arXiv:2404.08144.

Federal Bureau of Investigation (2025) *Internet crime report 2024.*

Festor, P., Habli, I., Jia, Y., Gordon, A., Faisal, A. A., & Komorowski, M. (2021) Levels of Autonomy and Safety Assurance for AI-Based Clinical Decision Systems, *Computer Safety, Reliability, and Security* (Vol. 12853, pp. 291–296), Springer International Publishing.

Gan, T., & Sun, Q. (2025) RAG-MCP: Mitigating Prompt Bloat in LLM Tool Selection via Retrieval-Augmented Generation, arXiv preprint arXiv:2505.03275.

Goecks, V. G., & Waytowich, N. (2024) COA-GPT: Generative Pre-Trained Transformers for Accelerated Course of Action Development in Military Operations, *2024 International Conference on Military Communication and Information Systems (ICMCIS)*, 01–10.

Goecks, V. G., & Waytowich, N. R. (2023) DisasterResponseGPT: Large Language Models for Accelerated Plan of Action Development in Disaster Response Scenarios, arXiv preprint arXiv:2306.17271.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023) Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection, arXiv preprint arXiv:2302.12173.

Hersey, P., Blanchard, K. H., & Johnson, D. E. (2013) *Management of organizational behavior: Leading human resources (10th edition)*, Pearson.

Johnson, M., & Bradshaw, J. M. (2021) How Interdependence Explains the World of Teamwork, *Engineering Artificially Intelligent Systems* (Vol. 13000, pp. 122–146), Springer International Publishing.

Kereopa-Yorke, B. (2024) ClausewitzGPT Framework: A New Frontier in Theoretical Large Language Model-Enhanced Information Operations, *Journal of Information Warfare*, 23(2).

LangChain. (2025) "Context Engineering", [online], https://blog.langchain.com/context-engineering-for-agents/.

MacCallum, N., & Lee, J. (2025) "OpenAI Cookbook: GPT-4.1 Prompting Guide", [online], https://cookbook.openai.com/examples/gpt4-1_prompting_guide.

Mbaiossoum, B. L., Mahamat, A. D., Batouma, N., Dionlar, L., Apollinaire, B. B., & Adam, I. O. (2025) How to Choose the Best AI LLM: A Guide to Navigating the Diversity of Models, *Journal of Information Systems Engineering and Management*, 10(34s), 221–232.

Microsoft AI Red Team. (2024) "Python Risk Identification Tool for generative AI (PyRIT)", [online], https://azure.github.io/PyRIT/.

NATO (2020) *Allied Joint Doctrine for Cyberspace Operations (AJP-3.20).*

NATO Cooperative Cyber Defence Centre of Excellence (2025) "Locked Shields 2025 Showcased Nations' Commitment to Defending Cyberspace", [online], https://ccdcoe.org/news/2025/locked-shields-2025-showcased-nations-commitment-to-defending-cyberspace/.

OpenAI (2025a) "Gpt-oss-120b & gpt-oss-20b Model Card", [online], https://openai.com/index/gpt-oss-model-card/.

OpenAI (2025b) "Introducing GPT-5", [online], https://openai.com/index/introducing-gpt-5/.

Pajo, P. (2025) Context Engineering: Enhancing Large Language Model Performance Through Comprehensive Contextual Management, Pre-print.

Qi, X., Huang, K., Panda, A., Henderson, P., Wang, M., & Mittal, P. (2023) Visual Adversarial Examples Jailbreak Aligned Large Language Models, arXiv preprint arXiv:2306.13213.

Sarjakivi, P. (2025) Artificial Intelligence Accelerated Cyber Operations: A Systematic Literature Review, *Journal of Information Warfare*, 24(1).

Sarjakivi, P., Ihanus, J., & Moilanen, P. (2025) Demonstration and Evaluation of Defensive Cyber Operations Decision-Making Model, *European Conference on Cyber Warfare and Security*, 24(1).

Smeets, M. (2022) The Role of Military Cyber Exercises: A Case Study of Locked Shields, *14th International Conference on Cyber Conflict*, 700, 9–25.

Snow, J. J. (2015) Entering the matrix: The challenge of regulating Radical Leveling Technologies*, Monterey, California: Naval Postgraduate School*.

SPLX (2025), "About Us", [online], https://splx.ai/about-us.

Vellum (2025) "LLM Leaderboard", [online], https://www.vellum.ai/llm-leaderboard.

Wade, N. M. (2023) *BSS7: The Battle Staff SMARTbook (7th ed.)*, The Lightning Press.

Wan, A., Wallace, E., Shen, S., & Klein, D. (2023) Poisoning Language Models During Instruction Tuning, arXiv preprint arXiv:2305.00944.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023) Jailbroken: How Does LLM Safety Training Fail?, arXiv preprint arXiv:2307.02483.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, arXiv preprint arXiv:2201.11903.

Willison, S. (2025) "The lethal trifecta for AI agents: Private data, untrusted content, and external communication", [online], https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/.

Yan, W. (2025) "Don't Build Multi-Agents, Cognition", [online], https://cognition.ai/blog/dont-build-multi-agents.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023) ReAct: Synergizing Reasoning and Acting in Language Models, arXiv preprint arXiv:2210.03629.

Yin, R. K. (2003) *Case study research: Design and methods (3rd ed)*, Sage Publications.

Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J. Z., & Fredrikson, M. (2023) Universal and Transferable Adversarial Attacks on Aligned Language Models, arXiv preprint arXiv:2307.15043.

VAOS: Vulnerability Attribute Ontology Score Framework for Evaluating Vulnerability Databases

Johnny Shaieb and John Hale
University of Tulsa, OK, USA

jas4252@utulsa.edu
john-hale@utulsa.edu

Abstract: Vulnerability repositories play a foundational role in enabling organizations to perform vulnerability scoring, prioritization, and threat modeling; however, they vary widely in the vulnerability attributes they define, require, and make available within their repository schemas. Although existing taxonomies describe numerous vulnerability characteristics, limited research evaluates how effectively real-world repositories support these practical security activities. This paper introduces the **Vulnerability Attribute Ontology Score (VAOS)**, a framework for evaluating vulnerability repositories rather than individual vulnerabilities. VAOS defines nineteen weighted attributes organized across mandatory, recommended, and optional tiers. The framework is applied to ten vulnerability repositories spanning more than five decades, revealing substantial variation in attribute coverage—particularly in contextual attributes—and demonstrating VAOS's value for repository evaluation, selection, and integration.

Keywords: Vulnerability database attribute ontology analysis

1. Introduction

1.1 Challenge

Vulnerability repositories underpin the daily workflows of modern security programs by enriching vulnerability scoring, prioritization, and threat modeling across tools and teams. Yet these repositories differ markedly in the attributes they define, require, and expose within their schemas. Although most repositories interoperate through shared identifiers such as **CVE** (and related classification schemes such as **CWE**), variability in attributive structure and availability can degrade the fidelity of scoring, prioritization, and attack modeling decisions. As a result, practitioners lack a concise, evidence-based method for comparing repositories based on the **presence and structure** of the attributes that matter operationally.

1.2 Gap

Although decades of work on taxonomies and ontologies have advanced how individual vulnerabilities are represented—for example, standardized identifiers such as CVE, classification schemes such as CWE, and severity scoring via CVSS—the literature provides little guidance for evaluating repository schemas themselves. In particular, there is no clear method for assessing which attributes repositories define and require, or how well those attributes collectively support prioritization and threat-informed analysis. Existing efforts often assume the source repository is sufficiently expressive; they do not offer a comparative approach to evaluate attribute coverage across real-world repositories, eras, or platforms.

1.3 Objective

In this paper, we present the Vulnerability Attribute Ontology Score (VAOS), a weighted framework grounded in ontological principles that is designed to assess vulnerability repositories, rather than individual vulnerability records. Within VAOS, the unit of analysis is defined at the level of repository schemas and their publicly documented metadata fields. It operationalizes a set of **nineteen attributes** organized into **mandatory, recommended, and optional** tiers, with weights reflecting historical prevalence and relevance to modern security operations. Critically, VAOS measures attribute presence and tier coverage; it does not evaluate the accuracy, integrity, or quality of a specific vulnerability within a repository.

1.4 Scope

This study evaluates vulnerability repository schemas by examining whether attributes are defined, required, or made available for use, and how those attributes map to the VAOS tiers and weights. It does not audit record-level vulnerability data for correctness. This scope ensures that the results reflect structural capability, rather than data hygiene or curation performance.

1.5 Contributions

This work makes following contributions:

- It defines a formal attribute ontology and weighting scheme (VAOS) that captures what repositories should record to support interoperability, automation, and threat-informed prioritization.
- It applies VAOS to ten historical and contemporary repositories spanning over five decades, producing a comparative analysis of attribute coverage and highlighting notable variation—especially in contextual fields (e.g., exploit maturity, access vector, authentication, classification, and compromise scope).
- It provides actionable guidance for repository evaluation, selection, and integration, enabling organizations to reason about repository fitness on the basis of attribute schemas rather than anecdotal feature lists.

1.6 Organization

Section 2 discuss prior taxonomy- and ontology-based approaches to vulnerability representation and motivates the need for a **repository-level** evaluation method. Section 3 defines the VAOS attributes, tiers, and weighting rationale. Section 4 applies VAOS to ten repositories and reports comparative findings across eras, including differences observed in contextual attributes that affect operational analysis. Section 5 concludes with implications for repository evaluation, selection, and integration, and outlines opportunities for extending VAOS to additional repositories and attribute families.

2. Vulnerabilities Taxonomy and Ontology Studies

2.1 Background and Related Work

Awareness of computer vulnerabilities predates the Internet itself. Early work, such as Ware's 1967 analysis of security and privacy in computer systems, emphasized the need for secure design and systematic reasoning about unauthorized access (Ware, 1967). This perspective was operationalized in the early 1970s when Saltzer introduced a structured list of vulnerabilities for the MULTICS operating system, identifying core attributes such as name, category, description, remedy, and status (Saltzer, 1973). As vulnerabilities increased in scale and complexity, numerous repositories emerged, often lacking standardized identifiers or consistent attribute structures, leading to fragmentation and duplication (Baker et al., 1999; Mann and Christey, 1999). These challenges motivated later efforts to formalize vulnerability representation through taxonomies (Abbott et al.) and ontologies (Akbar et al., 2023), including identifier schemes, weakness classifications, and scoring systems. While these approaches significantly improved how individual vulnerabilities are described and shared, they do not provide a method for evaluating the attribute schemas of vulnerability repositories themselves. This limitation motivates the need for a repository-level framework such as VAOS.

2.2 Vulnerability Taxonomy Studies

Vulnerability taxonomies provide hierarchical labels for organizing security flaws based on shared characteristics, enabling consistent classification and analysis. In cybersecurity, parent–child relationships (e.g., Injection → SQL Injection) support both human reasoning and machine processing and formed the basis of early vulnerability repositories. Seminal works such as Abbott et al. (1976) and Aslam (1995) exemplify early taxonomic approaches to vulnerability classification.

Within this model, higher-level categories capture broad classes of flaws, while child categories represent more specific manifestations. For example, "Buffer Overflow" may serve as a parent category with child categories such as "Heap-Based Buffer Overflow" and "Stack-Based Buffer Overflow," while "Injection" encompasses techniques such as SQL Injection and Cross-Site Scripting (XSS). This hierarchical structure enables precise categorization and supports systematic analysis across vulnerability databases.

2.2.1 The RISOS project

In April 1976, the National Bureau of Standards (NBS) issued a seminal report introducing one of the earliest taxonomies for classifying operating-system vulnerabilities, organized around five attributes: Name, Category, Description, Remedy, and Status (Abbott et al., 1976). The RISOS Project (Research into Secure Operating Systems) also identified fundamental vulnerability classes, including Incomplete Parameter Validation, Inconsistent Parameter Validation, Implicit Sharing of Privileged or Confidential Data, Asynchronous Validation/Inadequate Serialization, Inadequate Identification/Authentication/Authorization, Violable Prohibition or Limit, and Exploitable Logic Error. Within VAOS, the RISOS attributes Name, Category, Description, and Remedy map directly to Name, Category, Description, and Solution, enabling consistent alignment across historical and modern repositories.

2.2.2 A taxonomy of security faults in the UNIX operating system

Aslam (1995) proposed a vulnerability database schema that introduced several foundational attributes, many of which anticipated challenges later addressed by CVE. A key contribution was the use of a Name/Identification Number as the primary key for each vulnerability, enabling enumeration and alignment of reports from multiple sources. The schema also included fields such as Source, Description, Detection Technique, Classification, Workaround, Description of Fix, Literature, and Consequence.

The inclusion of a Literature field allowed multiple references to be associated with a single vulnerability entry, addressing fragmented reporting across disparate systems and anticipating CVE's objective of unifying vulnerability disclosures. Aslam's Classification attribute functioned as a taxonomy label, reflecting concepts later formalized as "levels of abstraction" (Baker et al., 1999) and contributing to the foundation of the Common Weakness Enumeration (CWE) framework.

Within VAOS, the Aslam attributes Name/Identification Number, Source, Detection Technique, Classification, Literature, and Consequence map directly to Unique Identifier, Reporting Source, Detection Source, Classification, Reference Source, and Consequence, respectively, enabling consistent schema alignment across historical and modern repositories.

2.3 Vulnerability Ontology Studies

In cybersecurity, ontologies provide structured, machine-readable models for representing vulnerabilities and their relationships to affected products, attack techniques, and defensive measures. Breitman et al. (2007) define an ontology as "a formal, explicit specification of a shared conceptualization," emphasizing standardized structure and shared meaning across systems. Applied to vulnerability data, ontologies enable multi-dimensional representations that support semantic interoperability and contextual analysis.

Ontology-based approaches allow vulnerabilities to be linked across related frameworks—for example, associating a CVE entry with its corresponding CWE weakness, relevant ATT&CK techniques, and D3FEND countermeasure—providing a more holistic view of the threat landscape (Akbar et al., 2023). Guo and Wang (2009) demonstrated how such semantic enrichment enables machine-understandable vulnerability representations, addressing limitations of purely syntactic descriptions and supporting automated security management.

Similarly, the Vulnerability Description Ontology (VDO) defines a comprehensive set of attributes to support automated vulnerability analysis across diverse systems and environments (Booth and Turner, 2016). Building on these principles, VAOS applies an ontology-informed approach to evaluate vulnerability repositories by assessing the presence and structure of attributes—such as identifier, weakness type, attack technique, access vector, and solution—rather than individual vulnerability records.

2.3.1 An ontology-based approach to model common vulnerabilities

Guo and Wang (2009) propose an ontology-based approach to model CVE entries, enabling machine-understandable representations for use in automated security systems. They argue that traditional CVE descriptions are largely syntactic, indicating the existence of a vulnerability while providing limited semantic detail about affected products or consequences. This limitation constrains automation in security processes such as those envisioned by the Security Content Automation Protocol (SCAP), motivating the need for semantic enrichment.

To address this gap, Guo and Wang apply the Web Ontology Language (OWL) to structure vulnerabilities around key elements, including identifiers, affected software, attack types, and potential impacts, while aligning with standards such as CVE, CWE, and CVSS. Related efforts, such as the Cybersecurity Vulnerability Ontology (CVO) (Syed, 2020), further demonstrate how ontologies can integrate structured and unstructured vulnerability data for real-time alerting and threat intelligence.

Within VAOS, attributes such as IT_Product, IT_Vendor, Attack_Method, and Attacker map directly to Product, Version, Attack Technique, and Access Vector, enabling semantic reasoning, improved interoperability, and enhanced automation in vulnerability assessment and decision-making.

2.3.2 Vulnerability description ontology (VDO)

Booth and Turner's Vulnerability Description Ontology (VDO) represents one of the most comprehensive ontology-based efforts to formalize vulnerability representation for automated analysis (Booth and Turner,

2016). VDO was explicitly designed to improve machine interpretability by decomposing vulnerabilities into a structured set of thirty ontology attributes, referred to as *noun groups*. Each noun group defines not only the attribute name and definition, but also its intended usage, allowed values, and relationships to other attributes.

A distinguishing feature of VDO is its close alignment with the Common Vulnerability Scoring System (CVSS). Rather than treating severity scoring as an external artifact, VDO embeds CVSS-related concepts directly into the ontology, enabling automated reasoning about exploitability, impact, and access requirements. This approach significantly advances semantic vulnerability representation compared to earlier syntactic descriptions.

However, VDO's unit of analysis is the individual vulnerability record, and its primary objective is to enhance automated scoring and analysis *within* a repository. It does not provide a method for evaluating or comparing vulnerability repository schemas themselves—that is, which attributes repositories define, require, or expose at a structural level. As a result, VDO does not address a key practical problem faced by modern security programs: determining whether a given vulnerability repository is *fit for purpose* with respect to prioritization, interoperability, and threat-informed analysis.

VAOS builds directly on VDO's ontological principles but shifts the unit of analysis from individual vulnerabilities to repository-level attribute schemas. In particular, VAOS adopts VDO's concept of attribute *usage* to categorize vulnerability attributes and assigns each attribute a weight according to its importance: **Mandatory (1), Recommended (0.6), or Optional (0.3)**. These weights reflect both historical prevalence in vulnerability repositories and relevance to contemporary security operations.

Rather than modeling how vulnerabilities should be described, VAOS evaluates whether repositories provide the attributes necessary to support modern security activities, including scoring, prioritization, interoperability, and threat-informed analysis. This distinction is critical: VDO improves *how* vulnerabilities are represented, while VAOS evaluates *whether* repositories are structurally capable of supporting analysis at scale. The following section formalizes this repository-level ontology and defines the full VAOS attribute set and weighting model.

2.4 Attributes

This subsection defines each attribute in detail, explaining its role in vulnerability characterization and justifying its assigned weight within the ontology.

Unique identifier is a unique value assigned to distinctly identify a specific vulnerability. A well-known example is the CVE number, widely used by vulnerability scanners. A typical format looks like **CVE-YYYY-NNNNN**. The "**YYYY"** presents the year the vulnerability was *assigned* (not discovered or published). The "**NNNNN"** provides a unique number assigned to the vulnerability for that year. This attribute remains essential for enabling traceability, cross-referencing, and consistent communication across tools, vendors, and researchers.

- Weight: Mandatory (1)

Vulnerability name references the affected operating system or application, describing the method of exploitation, or may include a commonly known nickname (e.g., EternalBlue). Its primary purpose is to facilitate easier recognition, communication, and contextual understanding of the vulnerability across technical and non-technical audiences.

- Weight: Mandatory (1)

Vulnerability description is a key attribute used to distinguish one security flaw from another. It should contain enough detail to allow a security expert to differentiate between vulnerabilities that may share similar names (Baker et al., 1999). (Aslam, 1995) notes that the description should clearly explain how the vulnerability can be exploited.

In addition to identification and exploitation, a well-written description supports traceability and effective communication across security platforms. It helps analysts evaluate the impact, understand the affected components, and determine appropriate remediation steps.

- Weight: Mandatory (1)

The **consequence** of a vulnerability refers to the effects that occur when a threat actor successfully exploits it, e.g. unauthorized access, data alteration or loss, arbitrary code execution, and the disclosure of sensitive information. Accurately describing these consequences is essential not only for understanding the nature of the flaw but also for assessing its severity and guiding remediation priorities.

- Weight: Mandatory (1)

The **reference source** identifies the origin of information about a vulnerability. These can include vendor advisories, security bulletins, exploit repositories, academic papers, and formal frameworks. It is a required attribute because reconciling multiple sources for the same vulnerability was one of the original challenges addressed by Dr. Dave Mann and Steve Christey in establishing CVE (Mann and Christey, 1999).

- Weight: Mandatory (1)

The **reporting source** attribute refers to the individual or organization that first discloses a vulnerability, such as a researcher, vendor, or CVE Numbering Authority (CNA).

- Weight: Mandatory (1)

The **solution** attribute refers to the corrective actions required to mitigate or eliminate a vulnerability. These actions may include applying patches, modifying configurations, disabling services, removing unauthorized users, or implementing compensating controls to mitigate the risk.

- Weight: Mandatory (1)

The **exploit maturity** attribute is adopted from the CVSS temporal exploitability metric originally defined in version 1 and is further developed to allow for a database to determine the maturity level of an exploit in relation to accessibility and weaponized usage. Prior to 2004, database attributes will be evaluated for the inclusion of exploit related references. The exploit maturity attribute includes the following categories:

Category	Description
None	No known exploit exists publicly or privately.
Proof of Concept	A technical demonstration exists showing the vulnerability can be exploited, but without weaponization.
Public Exploit	Fully functional exploit code is available publicly in repositories (Exploit-DB) or frameworks (Metasploit).
Weaponization	The exploit is actively used in malware, ransomware, or APT campaigns.

- Weight: Recommended (.6)

The **publish date** attribute is formally defined as the date on which a vulnerability was *first disclosed to the public*. While this date can often be inferred from reference and reporting sources, its significance was later formalized by MITRE. In 2016, MITRE introduced the optional field DATE_PUBLIC in the CVE JSON schema to explicitly capture this information (CVE Project, 2016).

Although not mandatory, the publish date plays a critical role in understanding the timeline of disclosure, assessing exposure windows, and correlating vulnerability data across platforms. It also supports historical analysis and helps distinguish between reserved, assigned, and disclosed states within the CVE lifecycle.

- Weight: Recommended (.6)

A **severity score** is a standardized numerical value that maps to severity categories (e.g., Critical, High, Medium, Low, None), representing the inherent seriousness of a security vulnerability. Severity scores focus solely on the technical impact of exploitation—such as data compromise, system disruption, or privilege escalation—and do not account for risk or likelihood of exploitation. The most widely adopted framework for calculating severity is the Common Vulnerability Scoring System (CVSS), which translates qualitative assessments into quantitative values on a scale from 0.0 (None) to 10.0 (Critical) (FIRST, 2015). On December 29, 2006, the "Common Vulnerability Scoring System" was discussed in IEEE Security & Privacy as being part of the National Vulnerability Database (NVD) ecosystem (Mell et al., 2006).

- Weight: Recommended (.6)

The **product** attribute refers to the affected system associated with the vulnerability. This can include, but is not limited to, hardware devices, operating systems, software applications, or configuration settings. To enable consistent identification of products affected by vulnerabilities such as those listed in CVE records, MITRE published the Common Platform Enumeration (CPE) specification. The framework defines a standardized, machine-readable format for naming applications, operating systems, and hardware platforms within enterprise environments (MITRE Corporation, 2007). In November 2009, CPE was integrated into the National Vulnerability Database (NVD) (Quinn et al., 2009). For vulnerabilities disclosed prior to 2007, the **Product** attribute will not

be evaluated against the CPE schema; instead, it will be reviewed to confirm the presence of a recognizable product name.

- Weight: Recommended (.6)

The **version** attribute identifies the specific release of a product affected by a vulnerability. This is important because different versions of the same product may contain varying features, configurations, or flaws. Including version information narrows the scope of exposure and supports more accurate remediation.

- **Weight:** Optional (.6)

The **access vector** attribute defines the level and type of access a threat actor must possess to exploit a given vulnerability. This can range from:

Category	Description
Physical Access	Hands on physical access to the device.
Logical Access	Authenticated access through a local user account.
Adjacent Access	Exploitation via shared local area networks (LANs), Bluetooth, or other proximity-based technologies.
Remote Access	Exploitation over broader networks like internet or wide area networks (WANs).

This classification is foundational to understanding the exposure surface of a vulnerability and plays a critical role in prioritization and remediation strategies.

- Weight: Recommended (.6)

The **authentication** attribute defines the type of authentication required to access a system (e.g., device, operating system, application) to exploit the vulnerability (NAIC, 2004). Within the CVSS version 2 base metrics, the Authentication metric includes the below specific categories (Mell et al., 2007):

Category	Description
None	Authentication is not required to exploit the vulnerability.
Single	The vulnerability requires the threat actor to be logged into the system to exploit the vulnerability.
Multiple	The vulnerability requires the threat actor to authentication two or more times to log into the system to exploit the vulnerability.

- Weight: Recommended (.3)

Similar to the consequence attribute, the **compromise scope** of a vulnerability refers to level of access achieved on the device, operating system, application or system by exploiting the vulnerability. The level of access can be generalized into the following categories:

Category	Description
Regular User	A basic user account with limited access to shared system resources.
Power User	A user account with elevated privileges beyond those of a regular user, but without full administrative rights.
Administrative User	A user account with full control over the system and its resources

- Weight: Optional (.3)

The **threat score** attribute is a quantitative metric designed to predict the real-world likelihood and potential impact of a vulnerability being exploited. Unlike traditional severity scores, such as the Common Vulnerability Scoring System (CVSS), threat scores incorporate dynamic and contextual data to reflect actual risk exposure.

Risk-based scoring models integrate multiple factors that include vulnerability weaponization, threat intel repositories, active exploitation, network location, asset contextualization, and environmental variables. This methodology enables organizations to prioritize vulnerabilities based on theoretical severity and on practical exploitability. These models represent a shift from subjective static severity ratings to intel driven threat intelligence indicators enabling more effective vulnerability remediation prioritization strategies.

- Weight: Optional (.3)

The **classification** attribute refers to the underlying flaw, whether in software or hardware, that enables a vulnerability to exist. It represents the level of abstraction used to describe the root cause of a security issue, ranging from low-level coding errors to architectural design deficiencies. This attribute empowers an organization to identify general hardware and software errors that contribute vulnerabilities, enabling more effective remediation strategies. In 2006, MITRE released a classification taxonomy called Common Weakness Enumeration (CWE), which is a list of hardware and software weaknesses (MITRE, 2022).

For vulnerabilities disclosed prior to 2006, the **classification** attribute will not be evaluated against the CWE schema; instead, it will be reviewed to confirm the presence of hardware or software weaknesses (e.g., buffer overflows, SQL Injection).

- Weight: Optional (.3)

The **Attack Technique** attribute describes how an adversary achieves a tactical goal against a vulnerable target using tactics and techniques (The MITRE Corporation, 2013). These actions may span the full intrusion lifecycle—from early reconnaissance (e.g., scanning, gathering organizational or personnel details) to initial access (e.g., exploiting default credentials, injection flaws, or phishing), and ultimately to executing code that escalates privileges or establishes persistence. The **Attack Technique** attribute is used to evaluate a vulnerability repository's ability to capture the adversary's tactics, technique, and procedures used to compromise a target.

- Weight: Optional (.3)

The **Detection Source** attribute refers to the unique identifier assigned by a specific platform or tool used to detect the presence of a vulnerability. These identifiers serve as internal references that facilitate tracking, correlation, and remediation across scanning platforms, exploit repositories, and penetration testing frameworks. The **Detection Source** attribute will be used to evaluate a repository's ability to reference platform-specific identifiers that support vulnerability detection, validation, and cross-platform correlation.

- Weight: Optional (.3)

2.5 Vulnerability Database Assessment

To illustrate the implementation of a VAOS assessment, a focused assessment is performed on "Fyodor's Playhouse" vulnerability database. Additionally, the findings from **ten** distinct databases have been consolidated into a single comparative table. This approach facilitates a streamlined presentation of key attributes while preserving the depth of insight across sources.

2.6 Individual Assessment

Fyodor's Playhouse, created by Gordon Lyon (known as Fyodor), is an early online archive of 419 exploit scripts collected between January 1996 and May 1999. This database was selected due to its historical significance, having been analyzed by Peter Mell—NVD founder—in (Mell, 1998) and (Mell, 1999).

A total **VAOS** score of **7.7** was achieved, reflecting coverage of 62% of all VAOS data attributes. This includes 71% of mandatory attributes, 67% of recommended attributes, and 20% of optional attributes.

VAOS Attribute	Weight	Tier	Fyodor's Playhouse	Fyodor's Attributes
Unique Identifier	1.0	Mandatory	0.0	
Vulnerability Name	1.0	Mandatory	1.0	Title
Vulnerability Description	1.0	Mandatory	1.0	Description
Consequence	1.0	Mandatory	1.0	Exploit & Full Info
Reference Source	1.0	Mandatory	1.0	Exploit & Full Info
Reporting Source	1.0	Mandatory	1.0	Author
Solution	1.0	Mandatory	0.0	
Exploit Maturity	0.6	Recommended	0.6	Exploit & Full Info
Publish Date	0.6	Recommended	0.6	Date
Severity Score	0.6	Recommended	0.0	
Product	0.6	Recommended	0.6	Vulnerable Systems
Version	0.6	Recommended	0.6	Vulnerable Systems

VAOS Attribute	Weight	Tier	Fyodor's Playhouse	Fyodor's Attributes
Access Vector	0.6	Recommended	0.0	
Authentication	0.3	Recommended	0.0	
Compromise Scope	0.3	Optional	0.0	
Threat Score	0.3	Optional	0.3	Compromise
Classification	0.3	Optional	0.0	
Attack Technique	0.3	Optional	0.0	
Detection Source	0.3	Optional	0.0	
Total	12.40	62%	7.7	
Mandatory	7.00	71%	5.0	
Recommended	3.60	67%	2.4	
Optional	1.50	20%	0.3	

2.7 Consolidated Assessment

Overall, the X-Force database achieved the highest VAOS score of 11.5, reflecting 90.55% attribute coverage. While CVE (2025) and NVD (2025) closely follow with scores of 11.2, the key differentiator lies in X-Force's inclusion of the compromise scope attribute, which was absent in both CVE and NVD. This additional attribute enhances X-Force's ability to contextualize the level of access gained through exploitation, contributing to its superior threat modeling capability.

Early-era databases—such as MULTICS (1972), BugTraq (1993), and VULDA (1996)—focused primarily on minimal attribute population, which offered limited contextual information. These repositories achieved an average VAOS attribute score of 5.6 and a z-score of -0.81, indicating they were significantly below the overall mean in terms of attribute completeness.

In contrast, mid-era databases like X-Force (1997), CVE (1999), and I-CAT (1999) began integrating more structured fields, including the foundational unique identifier (i.e., CVE) attribute that enabled interoperability across platforms. These databases achieved an average VAOS score of 8.2 and a z-score of 0.01, reflecting alignment with the overall mean and a 31% increase in the ability to contextualize vulnerabilities through structured attributes. Present-era databases like CVE (2025) and NVD (2025) drastically expanded their scope to capture high fidelity attributes (e.g., product, version, exploit maturity, access vector, authentication, classification) that helped them achieve a high VAOS score of 11.2 which is a 37% increase over the mid-era databases with an average z-score of .99. The product and version attribute fields enabled granular asset-based vulnerability correlation across enterprise environments. While the exploit maturity, access vector, authentication, and classification signaled the ability to leap from basic interoperability to future threat modeling.

Vuln Databases	Year	Rank	VAOS Score	VAOS Grade	Mandatory	Recommended	Optional	Z-Score
MULTICS	1972	8	5.5	43.31%	57.14%	28.57%	20.00%	-0.85
BugTraq	1993	9	4.4	34.65%	28.57%	42.86%	40.00%	-1.21
VulDA	1996	7	7	55.12%	57%	42.86%	80.00%	-0.37
X-Force	1997	1	11.5	90.55%	100%	92.86%	40.00%	1.09
DOVE	1998	5	9.6	75.59%	86%	71.43%	40.00%	0.48
Fyodor's Playhouse	1999	6	7.7	60.63%	71%	57.14%	20.00%	-0.14
CVE (1999)	1999	10	3	23.62%	43%	0.00%	0.00%	-1.67
I-CAT	1999	4	10.2	80.31%	86%	85.71%	40.00%	0.67
CVE (2025)	2005	2	11.2	88.19%	100%	92.86%	20.00%	1.00
NVD (2025)	2025	2	11.2	88.19%	100%	92.86%	20.00%	1.00
STDEV	3.08	**Mean**	8.13					

3. Conclusions

The **Vulnerability Attribute Ontology Score (VAOS)** offers a framework for evaluating the completeness and contextual capabilities of vulnerability databases. By illuminating how vulnerability classification has evolved from early taxonomies to modern ontologies, VAOS connects foundational ideas with today's needs for interoperability, automation, and threat modeling. The tiered weighting system—mandatory, recommended, and optional—reflects both legacy adoption and emerging approaches, enabling a nuanced assessment of database fidelity.

The review of **ten** databases shows clear differences in how well attributes are covered. Among them, IBM X-Force, CVE (2025), and NVD (2025) display the strongest levels of completeness. These results highlight the need to align vulnerability database attributes with changing security priorities, especially as organizations move toward risk-based decision-making and the use of real-time threat intelligence and modeling.

In summary, VAOS provides a practical framework for researchers, practitioners, and database architects to evaluate and improve vulnerability repositories. As the cybersecurity landscape evolves, frameworks like VAOS are essential in ensuring that vulnerability data remains interoperable, actionable, and contextualized to support remediation strategies.

Ethics statement: I have included an Ethics declaration at the end of the paper, before the references, which either states that ethical clearance was not required for the research or if it was, I have described the clearance obtained.

AI statement: AI was not used in creating the ideas or content of this paper. Standard Microsoft Word publishing and Grammarly formatting were used to correct misspellings, tense, grammar and sentence structure.

References

Abbott, R.P., Chin, J., Donnelley, J., Konigsford, W., Tokubo, S. and Webb, D. (1976) *Security analysis and enhancements of computer operating systems: The RISOS project final report (NBSIR 76-1041)*. National Bureau of Standards. Available at: https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nbsir76-1041.pdf.

Aslam, T. (1995) *A taxonomy of security faults in the UNIX operating system*. Master's thesis, Purdue University, CERIAS. Available at: http://ftp.cerias.purdue.edu/pub/papers/taimur-aslam/aslam-taxonomy-msthesis.pdf.

Akbar, K.A., Rahman, F.I., Singhal, A., Khan, L. and Thuraisingham, B. (2023) *The design and application of a unified ontology for cyber security*. National Institute of Standards and Technology. Available at: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=956387.

Baker, D.W., Christey, S.M., Hill, W.H. & Mann, D.E., *The Development of a Common Enumeration of Vulnerabilities and Exposures*, presented at the Second International Workshop on Recent Advances in Intrusion Detection, (7-9 September 1999). Available at: https://www.cve.org/Resources/Media/Archives/OldWebsite/docs/Development_of_CVE.html.

Booth, H. and Turner, C., (2016) *Vulnerability Description Ontology (VDO): A Framework for Characterizing Vulnerabilities*. NIST Internal Report (NISTIR) 8138. U.S. Department of Commerce, National Institute of Standards and Technology. Available at: https://csrc.nist.gov/files/pubs/ir/8138/ipd/docs/nistir_8138_draft.pdf.

Breitman, K.K., Casanova, M.A. and Truszkowski, W. (2007) *Semantic Web: Concepts, technologies and applications*. London: Springer-Verlag. Available at: https://doi.org/10.1007/978-1-84628-710-7.

CVE Project, (2016) *CVE JSON schema v2.0*. GitHub. Available at: https://github.com/CVEProject/cve-schema/blob/main/schema/archive/v2.0/JSON-file-format-v2.md.

Forum of Incident Response and Security Teams (FIRST) (2015) *Common Vulnerability Scoring System (CVSS) version 3.0: Specification document*. Available at: https://www.first.org/cvss/v3-0/specification-document.

Guo, M. and Wang, J.A. (2009) *An ontology-based approach to model common vulnerabilities and exposures in information security*. Proceedings of the 2009 ASEE Southeast Section Conference. Available at: https://sites.asee.org/se/wp-content/uploads/sites/56/2021/05/2009ASEESE034.pdf.

Mann, D.E. and Christey, S.M., (1999) *Towards a Common Enumeration of Vulnerabilities*. The MITRE Corporation. Available at: https://www.cve.org/Resources/General/Towards-a-Common-Enumeration-of-Vulnerabilities.pdf.

Mell, P.M., Scarfone, K.K. and Romanosky, S. (2006) *Common Vulnerability Scoring System*. IEEE Security & Privacy. Available at: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=50899.

Mell, P., Scarfone, K. & Romanosky, S., (2007) *A complete guide to the Common Vulnerability Scoring System (CVSS) version 2.0*. Forum of Incident Response and Security Teams (FIRST). Available at: https://www.first.org/cvss/v2/cvss-v2-guide.pdf.

MITRE Corporation, (2007) *Common Platform Enumeration (CPE): Name format and description (Version 1.0)* [PDF]. Available at: https://www.mitre.org/sites/default/files/pdf/07_0030.pdf.

MITRE, (2022) *CWE history*. Common Weakness Enumeration. Available at: https://cwe.mitre.org/about/history.html.

National Infrastructure Advisory Council (NAIC), (2004) *Common Vulnerability Scoring System (CVSS)*. Department of Homeland Security. Available at: https://www.first.org/cvss/v1/cvss-dhs-12-02-04.pdf.

Quinn, S., Waltermire, D., Johnson, C., Scarfone, K. & Banghart, J., (2009) *The technical specification for the Security Content Automation Protocol (SCAP): SCAP Version 1.0 (NIST Special Publication 800-126)*. National Institute of Standards and Technology. Available at: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-126.pdf.

Syed, R. (2020) *Cybersecurity vulnerability management: A conceptual ontology and cyber intelligence alert system*. *Information & Management*, 57(6). Available at: https://www.sciencedirect.com/science/article/abs/pii/S0378720620302718.

Saltzer, J.H., (1973) *Multics security holes list: Revision 5*. Massachusetts Institute of Technology. Available at: https://people.csail.mit.edu/saltzer/Multics/MHP-Saltzer-060508/filedrawers/030.multics-security-holes-1972-1975/Scan%201.PDF.

The MITRE Corporation (2013) *Enterprise techniques*. MITRE ATT&CK. Available at: https://attack.mitre.org/techniques/enterprise/.

Mell, P. (1998) Understanding the global attack toolkit using a database of dependent classifiers.

Mell, P. (1999) Understanding the world of your enemy with I-CAT (Internet-Categorization of Attacks Toolkit).

Ware, W.H. (1967) *Security and privacy in computer systems*. Santa Monica, CA: RAND Corporation. Paper presented at the Spring Joint Computer Conference, Atlantic City, NJ. Available at: https://nsarchive.gwu.edu/sites/default/files/documents/2828418/Document-01-Willis-H-Ware-RAND-Corporation-P.pdf.

Digital Forensic Readiness to Mitigate Insider Threats in the SaaS Cloud Environment

Gabriel Shoderu, Stacey Baror, Abiodun Modupe and Sheunesu Makura
University of Pretoria, South Africa

go.shoderu@gmail.com
stacey.baror@tuks.co.za
abiodun.modupe@cs.up.ac.za
makura.sm@up.ac.za

Abstract: Insider threats continue to pose significant risks in Software-as-a-Service (SaaS) environments, where legitimate users hold varying levels of access and control. Existing mitigation measures remain largely reactive, focusing on post-incident investigation and evidence recovery, which often result in delayed detection and incomplete forensics. A proactive and forensically sound approach is therefore required to identify and contain insider activity before major compromise occurs. This paper presents the Digital Forensic Readiness to Bust Insider Threats (DFR-BUST) model, a framework that embeds forensic readiness principles within SaaS environments to enable early detection, secure evidence capture, and legally defensible investigations. The model is aligned with the ISO/IEC 27043 digital investigation process, operationalising its readiness, acquisitive, and concurrent process classes. The model was evaluated using an experimental setup based on publicly available insider-threat datasets to demonstrate its readiness and detection capability. The evaluation confirmed that the proposed architecture supports proactive evidence generation, integrity verification, and traceable anomaly detection within a controlled environment. Unlike conventional reactive approaches, DFR-BUST provides a proactive, evidence-centric mechanism that enhances both detection accuracy and forensic admissibility. Its modular design ensures adaptability across cloud platforms while maintaining compliance with international forensic investigation standards. Overall, this work bridges the gap between intelligent analytics and digital forensic readiness. By ensuring that insider detection outputs are accompanied by verified, admissible evidence, the framework contributes a practical foundation for developing forensic-aware, cloud-based security systems.

Keywords: Insider threat, Cybercrime, Digital forensic readiness, SaaS cloud, Artificial intelligence, Anomaly detection

1. Introduction

Cloud computing has transformed how organisations store, process, and share data, with Software-as-a-Service (SaaS) platforms now underpinning most enterprise operations. This model offers scalability and accessibility but introduces new forensic and security complexities due to distributed control and shared responsibility between providers and clients. Within this environment, insider threats remain a persistent and costly risk, as authorised users, employees, contractors, or third-party administrators can misuse legitimate privileges to compromise data integrity, confidentiality, or availability (Greitzer et al, 2019; Sarhan and Altwaijry, 2023; Ponemon Institute, 2022).

The forensic challenge in SaaS environments lies in visibility and accountability. Multi-tenant architectures, delegated access rights, and rapid provisioning obscure traces of insider activity, while responsibility for evidence preservation is divided between provider and customer (Alenezi et al, 2020; CERT, 2016). Existing controls detect policy violations but rarely ensure that the digital traces they generate are complete, verifiable, and admissible. This highlights the need for Digital Forensic Readiness (DFR), a proactive capability that prepares cloud systems to capture and preserve evidence before incidents occur (Rowlingson, 2004; ISO/IEC 27043, 2015).

Most current cybersecurity mechanisms are reactive, initiating investigation only after an incident has occurred (Rowlingson, 2004). Such post-event analysis often yields incomplete or inadmissible evidence, delaying response and increasing cost. Recent studies on machine-learning-based anomaly detection (Tuor et al, 2017; Sarhan and Altwaijry, 2023) show promise for identifying insider behaviour, yet they overlook forensic soundness, traceability, integrity, and legal admissibility (Tan, 2001; ISO/IEC 27043, 2015). The present research aligns with ISO/IEC 27043, which defines readiness, acquisition, investigation, and concurrent process classes for digital investigations, extending its readiness class to cloud environments through an intelligent, forensic-aware design.

To address this gap, the study proposes the Digital Forensic Readiness to Bust Insider Threats (DFR-BUST) model, which embeds readiness, archival, and audit mechanisms within SaaS environments to ensure proactive detection and preservation of admissible evidence. The model operationalises ISO/IEC 27043 process classes, readiness, acquisitive, and concurrent, to create a seamless flow from evidence generation to analysis.

The study contributes in two ways:

- It introduces the DFR-BUST framework that integrates digital forensic readiness principles with intelligent analytics aligned to ISO/IEC 27043.
- It evaluates the framework experimentally using insider-threat data to demonstrate readiness effectiveness and evidence integrity.

The design of DFR-BUST is guided by four key quality attributes derived from software-architecture principles (Bass et al, 2021): reliability, ensuring consistent log and evidence capture; security, protecting against unauthorised access; auditability, guaranteeing traceability and admissibility; and cloud compliance, supporting scalability within SaaS ecosystems.

Having established the need for proactive and forensically sound insider-threat mitigation in SaaS environments, the next section explores prior research that underpin the proposed framework.

2. Background Literature

This section provides an overview of the key areas that inform this study. It begins with digital forensics and the evolution toward readiness, it then examines insider threats in Software-as-a-Service (SaaS) environments, and finally, the role of artificial intelligence in behavioural analytics. Together these areas provide the basis for developing the Digital Forensic Readiness to Bust Insider Threats (DFR-BUST) model.

2.1 Digital Forensics

Digital forensics is the discipline concerned with identifying, collecting, preserving, analysing, and presenting digital evidence in a scientifically reliable and legally admissible manner (Daniel and Daniel, 2011; Tan, 2001). Traditional investigations are reactive, beginning only after an incident has occurred, often yielding partial or tampered artefacts. As computing has shifted toward distributed and virtualised infrastructures, purely reactive approaches have proven insufficient (Rowlingson, 2004).

To address this limitation, research introduced Digital Forensic Readiness (DFR), the organisational capability to maximise evidential use while minimising investigation cost (Rowlingson, 2004). DFR emphasises preparation before incidents occur through structured logging, secure storage, and clear policies for evidence handling. The ISO/IEC 27043 standard formalises this concept by defining five investigation process classes: readiness, initialization, acquisitive, investigative, and concurrent. Readiness activities prepare systems for evidence capture, while concurrent activities ensure documentation and chain-of-custody integrity throughout.

Recent studies extended DFR to cloud environments, such as Kebande and Venter's (2017) Cloud Forensic Readiness as a Service (CFRaaS) model and Singh et al's (2022) SecureRS storage scheme. These studies highlight the need for automated, tamper-resistant evidence pipelines that maintain integrity and confidentiality. Yet DFR in SaaS contexts remains immature. Multi-tenant architectures and divided provider-customer responsibilities fragment evidence sources and hinder accountability (Alqahtany et al., 2019). Without a unified readiness process that coordinates collection and preservation across both parties, investigations risk data loss and unverifiable traces.

2.2 Insider Threats in SaaS Environments

Insider threats involve harmful actions by authorised users, employees, contractors, or partners, who misuse legitimate access (Greitzer et al., 2019). Behaviour may be intentional, such as data theft, or accidental, such as misconfiguring permissions (Ponemon Institute, 2022). Detection is difficult because insiders operate with valid credentials and knowledge of internal systems.

SaaS platforms amplify these challenges. Their web-based delivery allows access from many devices and locations, broadening the attack surface (Alenezi et al., 2020). Under the cloud's shared-responsibility model, providers secure infrastructure while customers manage user activity and configuration. This division often produces visibility gaps: providers hold infrastructure-level logs, whereas customers see only application-level events.

Audit trails, logins, data exports, and privilege changes, can reveal misuse if collected and preserved correctly. However, few organisations implement systematic, forensic-ready logging. As a result, evidence of intent or negligence may be incomplete or inadmissible during investigation. The Employee Management System (EMS) scenario later in this paper illustrates these risks where a user exporting bulk employee data after hours can only

be proven malicious if immutable audit logs exist. Hence, SaaS environments require frameworks that merge behavioural analytics with forensic readiness to ensure both early detection and evidential reliability.

2.3 Artificial Intelligence in Insider Threat Detection

Artificial intelligence (AI) and machine learning (ML) enable systems to analyse large behavioural datasets and flag anomalies indicative of insider misuse (Tuor et al, 2017; Sarhan and Altwaijry, 2023). Techniques include supervised classification, unsupervised anomaly detection, and deep learning. Feature-engineering methods such as Deep Feature Synthesis automatically derive higher-level behavioural indicators from raw logs (Kanter and Veeramachaneni, 2015). Because insider events are rare, many studies address class imbalance through resampling or cost-sensitive training.

While these approaches enhance detection accuracy, they seldom address forensic soundness. Metrics such as precision and recall dominate evaluation, whereas evidential integrity, chain-of-custody, and admissibility are often overlooked (Tan, 2001). Consequently, AI systems may identify anomalies yet fail to produce trustworthy evidence for investigators.

Recent literature therefore calls for integrating AI with DFR principles (Kebande and Venter, 2017; Singh et al, 2022). Such integration ensures that every analytical output is traceable to preserved source data and that alerts can serve as verifiable forensic artefacts. In this study, AI functions as a supporting component within a broader forensic-aware architecture rather than as the primary contribution. The goal is to combine intelligent detection with ISO/IEC 27043-compliant readiness processes so that insider-threat analytics yield both timely insights and admissible evidence.

Existing research demonstrates substantial progress in digital forensics, insider-threat analytics, and AI-based anomaly detection, yet these domains often evolve independently. Current cloud frameworks either focus on detection accuracy or on evidence preservation, seldom both. This fragmentation leaves SaaS organisations without a unified mechanism that ensures forensic readiness while supporting intelligent monitoring. Addressing this gap, the present work introduces the proposed DFR-BUST model, which consolidates readiness, detection, and preservation within a single ISO-aligned framework. The following section presents this model and the requirements guiding its design.

3. DFR-BUST Model

This section presents a holistic model that integrates digital forensic readiness with intelligent insider-threat monitoring in Software as a Service environment. The proposed model, DFR-BUST, prepares SaaS platforms for forensic investigation before incidents occur, ensuring that evidence generated during system operation remains authentic, traceable, and legally admissible (ISO/IEC 27043, 2015; Tan, 2001).

3.1 High level view of DFR-BUST Model

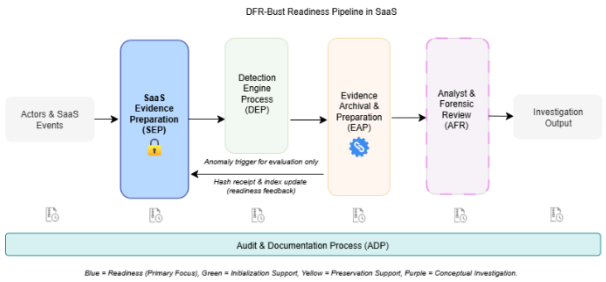


Figure 1: High-level view of the DFR-BUST readiness pipeline in a SaaS environment

Figure 1 illustrates the high-level pipeline. The SaaS Evidence Preparation (SEP) component performs pre-incident activities such as log generation, time synchronisation, normalisation, hashing, encryption, and daily archival. The Detection Engine Process (DEP) provides an initialization trigger by recognising anomalies and preserving the exact trigger context. The Evidence Archival and Preservation (EAP) component packages and verifies evidence, stores it in append-only repositories, and enforces access control (Kebande and Venter, 2017). The Analyst and Forensic Review (AFR) stage represents the conceptual hand-off to downstream examination and reporting (ISO/IEC, 2015). The Audit and Documentation Process (ADP) run across all components to maintain chain of custody, access logs, and hash receipts (Tan, 2001).

The study focuses on the readiness activities implemented in SEP, together with the preservation guarantees provided by EAP and the concurrent documentation maintained by ADP. DEP is included as a trigger and for evaluation of readiness effectiveness. AFR is included for completeness as a conceptual endpoint.

3.2 Mapping to ISO/IEC 27043

The ISO/IEC 27043 standard defines a comprehensive model for digital investigations through five process classes: readiness, initialization, acquisitive, investigative, and concurrent, each describing a stage in the life cycle of digital evidence (ISO/IEC 27043, 2015).

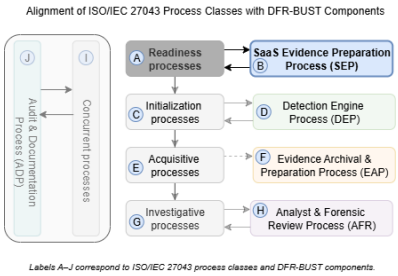


Figure 2: Alignment of ISO/IEC 27043 process classes with DFR-BUST components

To ensure methodological consistency and legal admissibility, the DFR-BUST model aligns its components with these process classes while maintaining a primary focus on forensic readiness. Figure 2 shows how each component supports a corresponding ISO process.

The SEP component directly implements the readiness class by enabling pre-incident log generation, normalization, and archival. DEP represents initialization by activating the investigation workflow once anomalies are detected. EAP performs acquisitive functions by packaging and verifying potential evidence. AFR embodies the investigative class by transforming preserved artefacts into meaningful insights. Finally, ADP operate as a concurrent layer that maintains chain of custody, provenance, and reproducibility across all stages.

This alignment demonstrates that while DFR-BUST supports the full investigative continuum, its main contribution lies in proactive readiness, ensuring that digital evidence is generated, protected, and auditable before incidents occur, creating a verifiable foundation for subsequent investigation.

Table 1: Mapping ISO 27043 to DFR-BUST Components

ISO 27043 Process Class	Primary Purpose	Corresponding DFR-BUST Component	Role in DFR-BUST Model	Evidence Outcome
Readiness	Pre-incident evidence preparation	SaaS Evidence Preparation (SEP)	Implements logging, normalisation, hashing, and archival	Structured, integrity-verified data
Initialization	Trigger and plan investigation	Detection Engine Process (DEP)	Generates anomaly-based alerts and preserves trigger context	Alert record and trigger dataset
Acquisitive	Collect and preserve evidence	Evidence Archival & Preservation (EAP)	Compiles, packages, and stores evidence immutably	Forensically packaged evidence
Investigative	Analyse and interpret data	Analyst & Forensic Review (AFR)	Examines evidence to establish intent or cause	Analytical report
Concurrent	Maintain documentation and oversight across all stages	Audit & Documentation Process (ADP)	Logs action, verifies chain of custody	Audit trail and verification receipts

3.3 Core Components of DFR-BUST Model

The DFR-BUST model comprises five coordinated components that collectively enable forensic readiness within a SaaS environment. SEP performs pre-incident collection and protection of activity logs. DEP recognises anomalies and initiates investigation triggers. EAP verifies and stores artefacts immutably. AFR supports human interpretation of preserved evidence, while the ADP maintain chain-of-custody and accountability across all stages.

Together, these components form a continuous readiness cycle in which evidence is captured, verified, and auditable before any incident occurs.

3.3.1 SaaS evidence preparation (SEP)

This component ensures that the SaaS environment produces logs that are structured, reliable, and legally admissible before incident occurs. It underpins forensic readiness by ensuring every action in the SaaS system can later be examined if necessary.

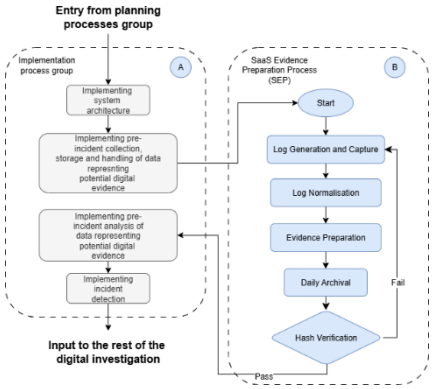


Figure 3: Integration of SEP process into the readiness process class of ISO/IEC 27043

The main activities include:

- Log Generation and Capture: Collects SaaS activities such as logon, logoff, device usage, file access, and web browsing in real time with synchronised timestamps.
- Log Normalisation: Standardises captured logs into a common format, removing duplicates and correcting inconsistencies.
- Evidence Preparation: Appends metadata such as department, user ID, and system identifiers; applies hashing and digital signatures to guarantee authenticity.
- Daily Archival: Batches and stores logs in secure repositories that preserve immutability and support later access.
- Hash Verification: After each archival cycle, verifies integrity through hash comparison. Successful checks confirm maintained readiness; failures trigger re-initialisation of log collection.

These steps ensure continuous evidence integrity and compliance with ISO/IEC 27043 readiness requirements. All generated logs and archives are concurrently registered in ADP for traceability.

3.3.2 Detection engine process (DEP)

This component represents the anomaly detection and behavioural analytics engine. It identifies deviations from behavioural baselines, flags potential insider threats, and initiates the investigative workflow.

Figure 4 illustrates the initialisation process in detail

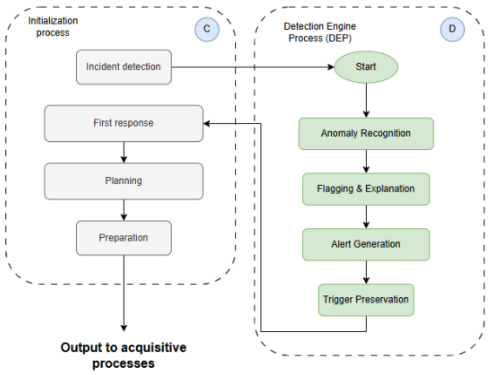


Figure 4: Integration of DEP within the initialization process class of ISO/IEC 27043

The main activities include:

- Anomaly Recognition: Applies detection models to SaaS logs and behavioural features, identifying patterns that deviate from established baselines.
- Flagging and Explanation: For each suspicious event, the system generates a flag accompanied by an explanation, such as unusual login time or abnormal data transfer.
- Alert Generation: Alerts are raised with structured contextual information, enabling investigators to understand the significance of the anomaly quickly.
- Trigger Preservation: Preserves datasets and metadata associated with each flagged anomaly to support further forensic analysis.

These activities ensure timely anomaly identification and systematic transition to investigation. The detection engine refines its baselines through adaptive learning while leaving validation and interpretation to analysts. Alert metadata and trigger details are automatically logged within ADP to maintain reproducibility.

3.3.3 Evidence archival and preservation process (EAP)

This component compiles, packages, and preserves all relevant data once an anomaly has been flagged, ensuring that contextual evidence remains complete, authentic, and admissible.

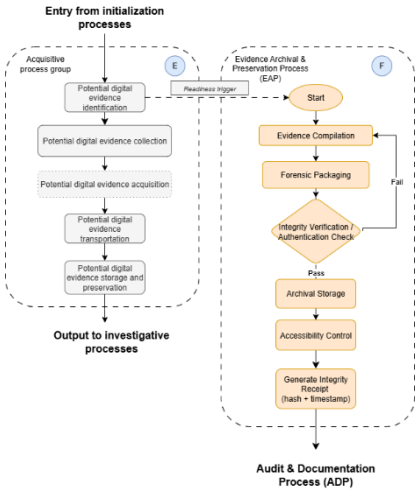


Figure 5: Integration of EAP within the acquisitive process class of ISO/IEC 27043

The main activities include:

- Evidence Compilation: Collects all relevant logs, engineered features, and contextual metadata surrounding a flagged anomaly, including before-and-after activity.
- Forensic Packaging: Bundles compiled information into a structured archive; applies hashing and digital signatures to confirm authenticity.
- Integrity Verification: Checks packaged evidence for alteration; failed checks trigger re-packaging before archival.
- Archival Storage: Stores verified evidence in immutable, access-controlled repositories.
- Accessibility Control: Monitors evidence accesses to preserve chain of custody.
- Integrity Receipt: Generates a hash and timestamp for every archived bundle and forwards them to ADP for verification.

These activities transform flagged anomalies into forensically reliable evidence bundles ready for investigation or legal review.

3.3.4 Analyst and forensic review process (AFR)

This component provides the human analytical layer where preserved evidence is examined, interpreted, and reported. Figure 6 illustrates the investigative process in detail.

The main activities include:

- Examination of Evidence: Validates whether anomalies flagged by the detection engine correspond to genuine insider behaviour.
- Contextual Analysis: Assesses evidence against insider-threat scenarios and organisational policies.

- Interpretation: Determines whether behaviour indicates malicious intent, negligence, or benign anomaly, ensuring that conclusions remain evidence-based and traceable.
- Reporting: Compiles findings into structured forensic reports, logged concurrently in ADP for audit tracking.

AFR bridges intelligent detection with human reasoning, ensuring that automated outputs are converted into defensible investigative insight.

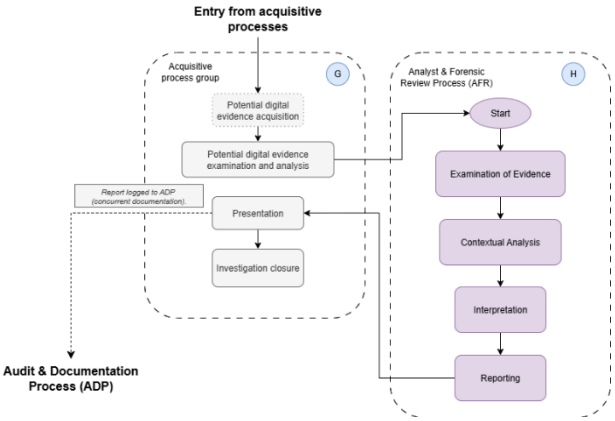


Figure 6: Integration of AFR within the investigative process class of ISO/IEC 27043

3.3.5 Audit and documentation process (ADP)

This component operates concurrently across all other processes to ensure documentation, reproducibility, and chain-of-custody assurance. The main activities include:

- Audit Trails: Automatically record every action from log capture to evidence analysis.
- Chain of Custody: Track and document access to evidence bundles.
- Reproducibility: Structure archived data and investigative steps so identical results can be reproduced.
- Quality Assurance: Conduct periodic integrity checks to confirm that forensic records remain complete and defensible.

As shown in Figure 1, the ADP runs concurrently beneath all components, reinforcing forensic integrity, traceability, and accountability at every stage of the digital investigation process. This concurrent layer ensures that DFR-BUST remains both operationally transparent and legally defensible.

The DFR-BUST model thus integrates the principles of ISO/IEC 27043 into a unified readiness-driven framework. Through its five components: SEP, DEP, EAP, AFR, and ADP, the model ensures that digital evidence in SaaS environments is captured, validated, and traceable before, during, and after incidents. While the framework spans all ISO process classes, its distinctive contribution lies in advancing forensic readiness, the capacity to generate reliable, auditable, and legally defensible evidence proactively.

This conceptual foundation also defines the logical structure upon which the intelligent forensic-readiness system is built. In the next section, DFR-BUST is operationalised as an adaptive system that embeds automated behavioural analytics, dynamic anomaly detection, and evidence-handling intelligence. This implementation translates readiness principles into active system behaviour, reinforcing the reliability, security, and auditability requirements identified earlier.

4. Intelligent Detection Engine

This section operationalises the Detection Engine Process (DEP) of the proposed DFR-BUST model. Machine-learning methods are embedded within this layer to detect behavioural deviations in a forensic-ready manner, ensuring that every anomaly flag remains traceable and admissible as digital evidence (Tan, 2001; ISO/IEC, 2015).

4.1 Dataset and Context

Experiments used the CERT r6.2 insider-threat dataset (Glasser and Lindauer, 2013), which extends earlier r4.2 and r6.1 versions to around 4,000 users across logon, device, file, and web logs. Unlike the older datasets, which contained approximately 1,000 users and multiple insiders per scenario, CERT r6.2 models a more realistic

enterprise with only five insider incidents, one per scenario. This design increases complexity and better reflects real-world conditions. Analysis focused on the Sales, Security, and Engineering departments containing these known insider cases, following some methodological precedents by Tuor et al (2017) and Sarhan and Altwaijry (2023).

4.2 Feature Engineering

Behavioural features were derived per user-day, representing actions such as after-hours logins, removable-drive use, file transfers, and device changes. Department-level aggregation reduced noise while preserving contextual meaning. Initially, features were manually selected based on domain knowledge, but model performance remained inconsistent until Deep Feature Synthesis (DFS) was incorporated. Following Kanter and Veeramachaneni (2015), DFS automatically generated higher-level features, such as frequency ratios, deviations, and temporal distributions, which were combined with manually selected ones to improve detection accuracy and behavioural representation within each department.

4.3 Models and Tuning

Two complementary anomaly-detection models were implemented within the DEP layer:

- One-Class SVM (OC-SVM): an unsupervised anomaly detector tuned via nu and contamination parameters to manage class imbalance. It was effective for identifying rare or subtle deviations.
- Isolation Forest (iForest): a tree-based ensemble method that isolates anomalies efficiently and is robust to noise and scaling issues.

Both models underwent iterative hyper-parameter tuning. PCA dimensionality reduction was tested for iForest to improve interpretability. Precision–recall balance was maintained through threshold and contamination adjustments. Together, the models form the analytical core of the DFR-BUST detection layer as shown in Figure 7, complementing forensic readiness by generating verifiable anomaly alerts.

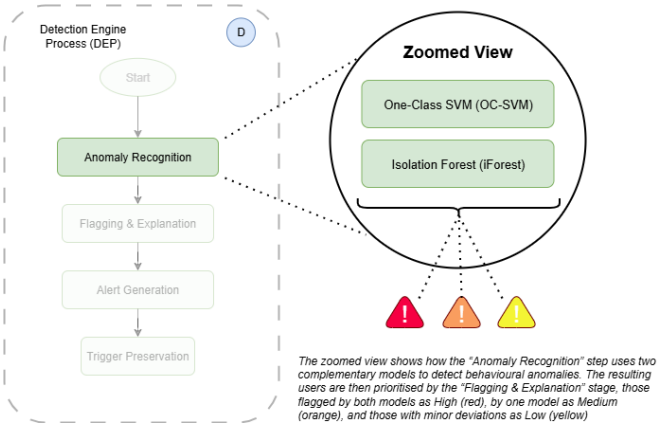


Figure 7: Integration of OC-SVM and iForest Models within the DEP Layer

4.4 Dataset Scenario 1: Model Performance Comparison

The first insider scenario depicts a user who began logging in after hours, using a removable drive, and uploading sensitive data externally before resignation. To benchmark results, Table 2 compares outcomes by Bin Sarhan and Altwaijry (2023) on the older CERT r4.2 dataset with results from this study using CERT r6.2.

Table 2: Comparison of Model Performance on Insider Scenario 1

Model	Dataset Version	Users	Insiders per Scenario	Precision	Recall	F1	Accuracy
OC-SVM (Bin Sarhan & Altwaijry, 2023)	CERT r4.2	~ 1,000	Multiple	0.94	0.86	0.89	0.86
iForest (Bin Sarhan & Altwaijry, 2023)	CERT r4.2	~ 1,000	Multiple	0.92	0.91	0.88	0.91
OC-SVM (This Study)	CERT r6.2	~ 4,000	1 per scenario	1.00	0.50	0.67	0969

Model	Dataset Version	Users	Insiders per Scenario	Precision	Recall	F1	Accuracy
iForest (This Study)	CERT r6.2	~ 4,000	1 per scenario	1.00	0.50	0.67	0.969

The expanded dataset and the single-insider structure naturally lowered recall but maintained high precision and accuracy, demonstrating practical forensic applicability rather than purely statistical optimisation.

4.5 All-Scenario Performance Summary

Performance across all five CERT r6.2 scenarios is summarised in Table 3. These include:

1. a user exfiltrating data before resignation
2. a user soliciting competitors while using removable drives
3. a user installing keyloggers and sending mass emails
4. a user accessing others' machines for file exfiltration
5. a user uploading company data to personal cloud storage

Table 3: Summary of Model Performance Across All Scenarios (CERT r6.2)

Scenario	Department	Model	Precision	Recall	F1	Accuracy
1	Sales	OC-SVM	1.000	0.500	0.667	0.969
1	Sales	iForest	1.000	0.500	0.667	0.969
2	Sales	OC-SVM	1.00	0.50	0.67	0969
2	Sales	iForest	1.000	1.000	1.000	1.000
3	Security	OC-SVM	0.333	1.00	0.500	0.935
3	Security	iForest	1.000	1.000	1.000	1.000
4	Engineering	OC-SVM	1.000	0.500	0.667	0.969
4	Engineering	iForest	0.182	1.000	0.308	0.719
5	Engineering	OC-SVM	0.000	0.000	0.000	0.938
5	Engineering	iForest	0.143	1.000	0.250	0.625

These results highlight the trade-off between model sensitivity and precision. iForest achieved perfect recall in Sales and Security scenarios, while OC-SVM offered higher selectivity. Together, they provide complementary strengths, balancing false-positive control with anomaly coverage. In addition, results indicate that departmental context significantly affects performance. For each department, the feature model pairing that best replicates its behavioural characteristics should be chosen, for example, engineering activity patterns differ greatly from those in sales or security. Therefore, the current combination performs well in some departments and could be replaced or fine-tuned in others to improve overall detection fidelity.

4.6 Flag Prioritisation and Forensic Integration

Within the DEP, outputs from both models were merged into a prioritised flagging system:

- High Priority: Users flagged by both OC-SVM and iForest.
- Medium Priority: Users flagged by one model but with high anomaly scores.
- Low Priority: Weak or inconsistent anomaly detections retained for reference.

Each flag is logged with model parameters and timestamps, then archived by ADP to ensure traceability and reproducibility. This process converts statistical anomaly detection into forensically admissible digital evidence.

5. Case Scenario of DFR-BUST Model

To illustrate the practical application of the proposed model, the following case scenario describes how the DFR-BUST framework operates within a cloud-based Employee Management System (EMS).

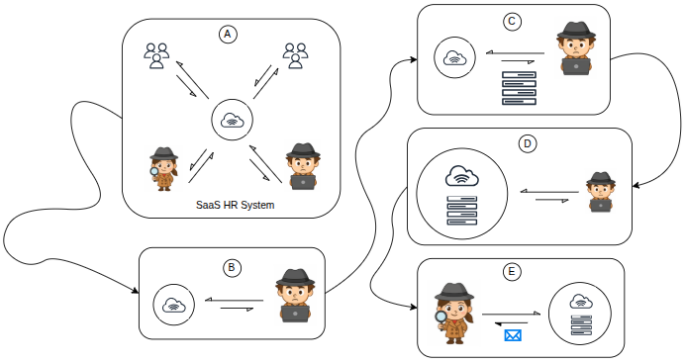


Figure 8: Case Scenario of DFR-BUST Model using Alice and Bob

At the readiness phase (A), the EMS continuously collects and prepares logs. A user (B) performs data-access activities outside business hours, triggering a deviation from baseline behaviour. The monitoring layer (C) captures detailed metadata, timestamps, file paths, and access volumes, and forwards it to the Detection Engine (D), where both OC-SVM and iForest models flag the event as anomalous. A structured alert is generated, and the Analyst and Forensic Review (E) process validates the findings, packaging the related data through the archival and audit layers.

This sequence demonstrates how DFR-BUST converts raw logs into actionable forensic intelligence while maintaining compliance and traceability. The model operationalises ISO/IEC 27043 by ensuring that evidence preparation, detection, preservation, analysis, and documentation function cohesively. Collectively, these processes uphold reliability, security, auditability, and cloud compliance, achieving the ISO goal of maintaining evidence that is both credible and legally defensible. The model's layered design confirms readiness through proactive preparation, traceable detection, and verifiable preservation.

6. Comparison with Related Work

Compared with prior studies, the DFR-BUST model offers a distinct integration of forensic readiness and intelligent anomaly detection. Tuor et al (2017) demonstrated that deep learning models could detect insider threats on early CERT datasets, but their approach remained performance-driven and did not address evidence admissibility or forensic soundness. Sarhan and Altwaijry (2023) explored hybrid supervised–unsupervised detection methods with strong precision metrics, yet without a readiness framework to preserve alerts for legal review. Similarly, Glasser and Lindauer (2013) introduced the CERT synthetic datasets to support insider research but did not propose a readiness model for operational environments.

In contrast, this study extends the forensic dimension by embedding AI models within the ISO/IEC 27043 process classes, ensuring that each anomaly flag is traceable, verifiable, and stored with chain-of-custody guarantees. The integration of Deep Feature Synthesis (DFS) and department-specific behavioural baselines provides contextual intelligence absent in prior work.

Moreover, while earlier studies emphasised accuracy over reliability, the proposed DFR-BUST model aligns with four critical dimensions. (Rowlingson, 2004; Tan, 2001; Greitzer et al, 2019; ISO/IEC 27043, 2015; Alqahtany et al, 2019)

- Reliability, as represented by the DFR principle, ensures that evidence remains consistent, complete, and scientifically defensible.
- Security, aligned with insider threat mitigation, focuses on proactively identifying and containing malicious activity.
- Auditability, rooted in forensic methodology, maintains an unbroken chain of custody and transparent documentation.
- Cloud Compliance, adherence to legal, operational, and privacy requirements across distributed SaaS infrastructures.

Collectively, these dimensions position DFR-BUST as both a technically adaptive and legally defensible advancement in insider-threat mitigation, bridging the gap between intelligent detection and forensic credibility in cloud environments.

7. Discussion and Conclusion

The findings confirm that embedding anomaly detection within a forensic-ready architecture enhances both detection reliability and evidential value. Although recall remains moderate, precision and admissibility are maintained, showing that AI-driven detection and forensic soundness are complementary rather than competing objectives (Tan, 2001; ISO/IEC 27043, 2015). The combined use of One-Class SVM and Isolation Forest balances sensitivity and precision while ensuring that all alerts remain traceable and legally defensible.

This study advances current practice by integrating Digital Forensic Readiness (DFR) and intelligent behavioural analytics within a unified, ISO-aligned framework. The proposed DFR-BUST model demonstrates that proactive evidence preparation, anomaly detection, archival, and audit documentation can operate cohesively in cloud environments, enabling organisations to generate verifiable and legally defensible digital evidence.

Some limitations remain. Detecting privileged or low-activity users (as in Scenario 5) was challenging due to limited behavioural signals. Computational constraints also restricted experimentation to selected departments, which may affect generalisability. Future implementations should therefore include richer telemetry, such as HR metadata and cross-departmental context, to improve detection coverage.

Further work should explore adaptive, multimodal learning that integrates enterprise telemetry with behavioural analytics while maintaining forensic traceability. Testing across additional datasets and live SaaS environments will help validate scalability and reliability. Addressing these aspects will enable DFR-BUST to mature into a robust, domain-independent framework that strengthens organisational resilience and supports trustworthy digital investigations.

Ethics Declaration: This study used only publicly available datasets and secondary scholarly sources. No human participants or personal data were involved; therefore, ethical clearance was not required.

AI Declaration: OpenAI's ChatGPT was used to assist with editing and summarizing reviewed materials for clarity and structure. All conceptual development, data analysis, experimental design, interpretation of results, and final decisions regarding structure, argumentation, and inclusion of content were made by the author, ensuring originality and accuracy.

References

Alenezi, M., Alarifi, A. and Alsaeedi, A. (2020) "Cloud forensic readiness: A framework for proactive digital investigations in SaaS environments", Journal of Cloud Computing, Vol. 9, No. 23, pp. 1–18.

Alqahtany, S., Clarke, N.L. and Furnell, S. (2019) "Developing a forensic readiness framework for cloud computing", Journal of Information Security and Applications, Vol. 45, pp. 1–11.

Bass, L., Clements, P. and Kazman, R. (2021) Software Architecture in Practice, Addison-Wesley, Boston.

Bin Sarhan, B. and Altwaijry, N. (2023) Insider Threat Detection Using Machine Learning Approach. Applied Sciences, 13 (1), 259. https://doi.org/10.3390/app13010259

CERT (2016) Common Sense Guide to Mitigating Insider Threats, 5th ed., Software Engineering Institute, Carnegie Mellon University, Pittsburgh.

Daniel, J. and Daniel, G. (2011) "Digital forensics: An overview of challenges in evidence collection and analysis", International Journal of Computer Applications, Vol. 36, No. 11, pp. 1–7.

Glasser, J. and Lindauer, B. (2013) "Bridging the gap: A pragmatic approach to generating insider threat data", Proceedings of the IEEE Security and Privacy Workshops, San Francisco, pp. 98–104.

Greitzer, F.L., Kangas, L.J. and Noonan, C.F. (2019) "A comprehensive framework for insider threat detection", Computers & Security, Vol. 87, pp. 101580–101590.

ISO/IEC (2015) Information Technology — Security Techniques — Incident Investigation Principles and Processes (ISO/IEC 27043:2015), International Organization for Standardization, Geneva.

Kanter, J.M. and Veeramachaneni, K. (2015) "Deep Feature Synthesis: Towards automating data science endeavours", IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, pp. 1–10.

Kebande, V.R. and Venter, H.S. (2017) "A cloud forensic readiness as a service model", Computers & Security, Vol. 70, pp. 773–785.

Ponemon Institute (2022) 2022 Cost of Insider Threats: Global Report, Ponemon Institute, Traverse City, MI.

Rowlingson, R. (2004) "A ten step process for forensic readiness", International Journal of Digital Evidence, Vol. 2, No. 3, pp. 1–28.

Sarhan, M. and Altwaijry, H. (2023) "A hybrid machine learning framework for insider threat detection", Expert Systems with Applications, Vol. 219, pp. 119634–119642.

Singh, A., Tripathi, M. and Kumar, S. (2022) "SecureRS: A secure cloud-based forensic-ready storage architecture", Future Generation Computer Systems, Vol. 129, pp. 214–229.

Tan, J. (2001) Forensic Readiness, @Stake, Cambridge, MA.

Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N. and Robinson, S. (2017) “Deep learning for unsupervised insider threat detection in structured cybersecurity data streams”, Proceedings of the IEEE Workshop on Cybersecurity and Machine Learning, Prague, pp. 1–7.

Social Media Misinformation in Europe and Africa: A Systematic Literature Review

Nokuthaba Siphambili, Errol Baloyi, Elekanyani Mukondeleli and Molebogeng Latakgomo
Council for Scientific and Industrial Research, Pretoria, South Africa

nsiphambili@csir.co.za
ebaloyi2@csir.co.za
emukondeleli@csir.co.za
mlatakgomo@csir.co.za

Abstract. In an era of rapid digital transformation, social media has emerged as a dominant source of information and is also used for misinformation. This creates a significant challenge because misinformation can influence public perceptions and behaviours related to online security. As a result, this paper compared how misinformation spreads on social media in Europe and Africa. Exploring the key factors influencing misinformation dynamics, the role of regulatory frameworks, and the effectiveness of fact-checking initiatives in both regions. It also assessed deepfakes' role in strengthening the spread of misinformation. A systematic literature review was used to gain these insights, which included exploring social media's role in spreading misinformation. The results highlighted the need for awareness and exposed the threat posed by deepfakes. Furthermore, the results also yielded that to tackle misinformation in this digital era, there is a need for a multi-stakeholder approach and cross-regional collaboration and use of deepfake detection tools to foster a more informed digital society.

Keywords: Misinformation, Social media, Disinformation, Cybersecurity, Africa, Europe

1. Introduction

Today, in this digital era, the world is experiencing a surge of manipulated information. This manipulated information is often referred to as misinformation; however, two other related terms are associated with manipulated information, which are fake news and disinformation. Misinformation is defined as false or misleading information distributed without intent to deceive (Aïmeur, Amri and Brassard, 2023). Fake news refers to misinformation presented as news articles. In addition, fake news is defined as incorrect or misleading information published as news, frequently to deceive (Aïmeur, Amri and Brassard, 2023). The key difference between misinformation and disinformation lies with the intent of the spreader. Misinformation is usually shared unintentionally, while disinformation is deliberately created to deceive, often targeting specific groups. Moreover, many researchers have noted that determining the intent behind the information is often challenging (Wu, Morstatter and Liu, 2019; Onigbinde and Oloyede, 2024). Therefore, this review uses "misinformation" as a broad term to encompass all instances of false or misleading information.

The spread of misinformation can often lead to serious consequences across various aspects of society, including politics, security, and the economy (De Jager, Marivate and Modupe, 2023). This may include a decline in trust in institutions, the spread of fear and panic, and, in some cases, the erosion of democracy (Onigbinde and Oloyede, 2024). However, misinformation is not a new phenomenon; it has been around for decades. Mahlasela et al (2024) in their investigation of the impact of artificial intelligence and machine learning on media manipulation, they noted that the creation and distribution of manipulated content will continue to rise due to the increasing accessibility of technology and the ease of spreading information through social media platforms. An observation that is also noted by (Aïmeur, Amri and Brassard, 2023).

Misinformation has evolved in many ways over the years (Wasserman and Madrid-Morales, 2019). In Europe, which is the focus of this study, research examining misinformation from various fact-checking organizations across different continents found that many European countries are frequently targeted by political misinformation, particularly regarding foreign politics (Cazzamatta, 2024). However, this has not always been the case. BBC News (2025) reported on a social media misinformation campaign that went viral, fuelling conspiracy theorists and social media speculation. This campaign involved the false claim of the passing of Queen Elizabeth II.

In Africa, which is also a focus of this study, misinformation campaigns have primarily been used to influence political agendas (Wasserman and Madrid-Morales, 2019). In South Africa, one of the countries in Africa, (Mare, Mabweazara and Moyo, 2019) [9] noted the use of bots in a misinformation campaign by the Guptas to divert attention from state capture. Chenzi (2021) highlighted how social media platforms are gradually becoming a key factor in South Africa's contemporary xenophobia challenge. Finally, more recently, as part of the Southern African Development Community (SADC) peacekeeping force, a misinformation campaign gained traction,

falsely claiming that South African National Defence Force (SANDF) soldiers were held captive or hostage by the M23 rebels (Van Tilburg, 2025).

A similar observation was noted during Russia's invasion of Ukraine, highlighting how disinformation can spread rapidly and have a significant impact, particularly through social media (Morosoli et al., 2025). During the invasion, Russia used this tactic to spread false claims, such as the Ukrainian president surrendering, Ukrainian soldiers laying down their arms and waving a white flag, and justifications for the invasion aimed at swaying public opinion and garnering support from other countries (Baloyi et al., 2024). This conflict underscored how, when combined with false information, social media can go beyond its original purpose as a communication tool and become a powerful weapon capable of shaping public perception, influencing opinions, and ultimately affecting real-world events.

The individuals behind misinformation campaigns vary. A study by Cazzamatta (2024) found that a significant portion of false information originates from unidentified or seemingly ordinary social media users. This was evident in the case of Queen Elizabeth II's rumoured death, where an ordinary account initially spread the false claim, which was later amplified by fake accounts (BBC News, 2025). However, misinformation is not always spread by individuals alone; in some cases, digital technologies such as bots, which are automated software designed to perform repetitive tasks online, play a crucial role in fuelling these campaigns (Mare, Mabweazara and Moyo, 2019). The extent to which social media contributes to misinformation warrants an investigation, particularly in Africa and Europe, which are the focus of this review.

2. Problem Statement

The main research problem is that the current understanding of how misinformation spreads on social media platforms in Europe and Africa is currently underexplored, which could potentially leave everyday social media users in both regions susceptible to misinformation campaigns. The following research objectives (RO) guided the study:

RO1: *To analyse the key factors influencing the spread of misinformation on social media platforms in Europe and Africa.*

RO2: *To identify similarities and differences in misinformation trends, propagation mechanisms, and audience susceptibility in Europe and Africa.*

RO3: *To assess the effectiveness of current misinformation detection and mitigation strategies across the two regions.*

R04: *To propose a framework for a more comprehensive approach to combating misinformation, considering regional differences and best practices.*

3. Methodology

A systematic literature review (SLR) is carried out to address the main research problem, which investigates existing literature on how social media influences misinformation in Europe and Africa. As indicated, the aforementioned research objectives guided this study. The Preferred Reporting Items for Systematic and Meta-analysis (PRISMA) framework was used to target publications between 2020 and 2025 with search terms ("misinformation" AND "social media" AND "Europe" AND "Africa" AND "cybersecurity") (Page et al., 2021). The following databases were used: Scopus, Elsevier, Google Scholar, and Taylor and Francis Online for comprehensive results.

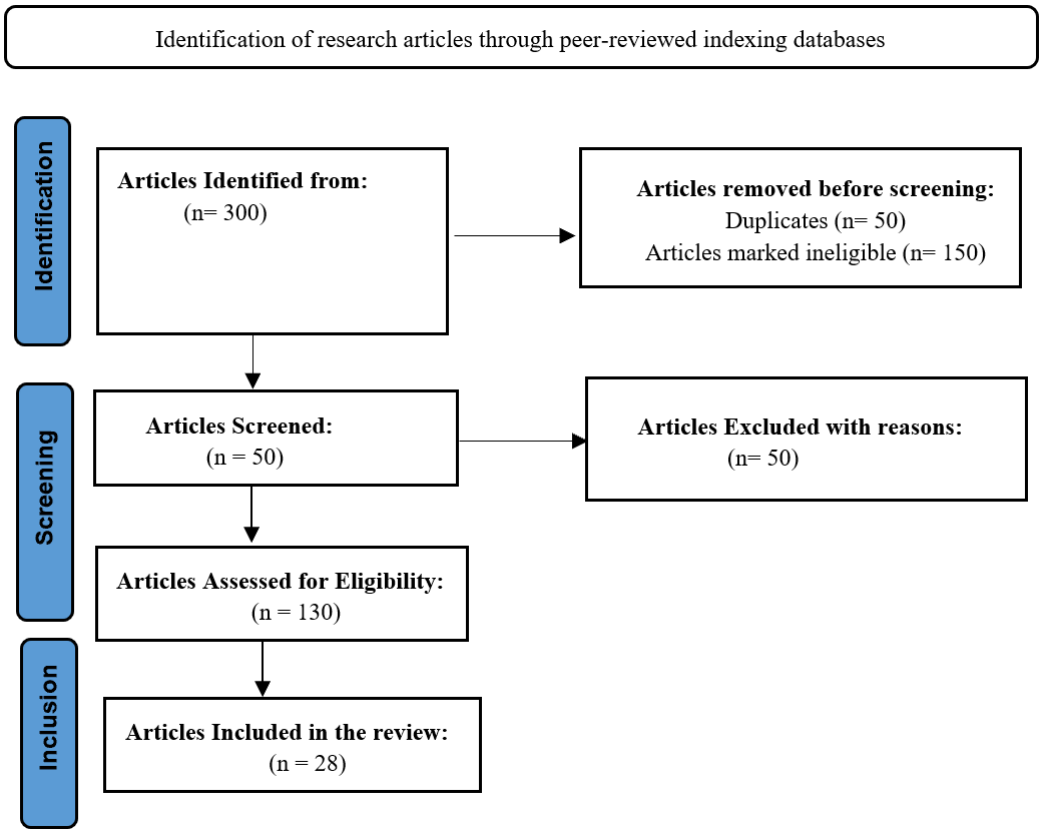


Figure 1: Representation of the PRISMA Framework

4. Literature Review

4.1 Factors Influencing the Spread of Misinformation

Social media algorithms prioritize engagement-driven content, often favoring sensational, emotionally charged, or polarizing information because such content tends to generate more interaction. This process may unintentionally spread misinformation. Although European platforms have taken steps to control material, there is still uneven enforcement of these regulations (European Commission, 2025). Effective content filtering in Africa is hindered by a lack of resources and issues with digital infrastructure (Mtsweni et al., 2023). Bolsonaro's usage of Telegram and YouTube is a prime example of how algorithms favor content that influences public opinion, even when it contains false information (Di Nubila et al., 2023). In this regard, social media algorithms reinforce divisions of opinion and spread false information by fostering the development of echo chambers and filter bubbles.

Political environments, including governmental structures, historical legacies, and media systems, significantly influence misinformation strategies. The rise of far-right populists like Guido Rios in Uruguay, Antonio Kast in Chile, Jair Bolsonaro in Brazil, and Javier Milei in Argentina demonstrates how these political leaders take advantage of political unrest to spread false information, particularly in areas with a history of authoritarianism and lax regulations (Kestler, 2022). Latin America's presidential systems and military dictatorship legacies influence the use of disinformation by both left- and right-wing groups, in contrast to Europe's corporatist and Mediterranean political systems (Kestler, 2022). Like Steve Bannon, leaders such as Santiago Abascal in Spain and Donald Trump in the United States have used disinformation as a weapon to attack democratic institutions and mainstream media (Hameleers and Minihold, 2022). While political media and state-sponsored disinformation are more prevalent in Europe, misinformation frequently spreads through informal networks like WhatsApp and local community discussions in Africa. This illustrates the various ways political contexts impact the dissemination of false information (Mtsweni et al., 2023).

4.2 Role of Social Media Platforms in Facilitating Misinformation

Social media platforms act as both enablers and mitigators of misinformation (Bhardwaj and Kumar, 2023). Platforms like Facebook, Twitter, and TikTok have difficulty enforcing standards to prevent false information,

especially in areas where influence on politics is common. Research shows that where political forces restrict regulation or where content moderation is weak, disinformation thrives (The Guardian, 2025). Social media platforms play a pivotal role in the dissemination of misinformation, influenced by their design, algorithms, and user interactions. Developing solutions to counteract false information on the internet requires an understanding of these systems (The Guardian, 2025; Bhardwaj and Kumar, 2023).

One important contributing component is algorithmic amplification, where social media sites employ algorithms that rank information according to user involvement (Mtsweni et al., 2023). This frequently leads to the promotion of emotionally charged or dramatic posts, which can accelerate the spread of false information. Algorithm amplifications lead to filter bubbles, where users are mostly exposed to material that supports their preconceived notions, increasing their susceptibility to false information. Due in large part to algorithmic amplification, studies have shown that false information is more likely to be shared and spread more quickly and extensively than accurate information (Fraga-Lamas and Fernandez-Carames, 2020). Falsehoods can spread quickly before fact-checking procedures take effect as users transmit content across large networks with a single click (Shu et al., 2020). Furthermore, multimedia manipulation and deepfakes have become particularly troublesome types of disinformation. The detection of misleading information is made more difficult by the fact that these alterations, which are frequently powered by AI technologies like deep-fake videos or GAN-generated graphics, are very hard for users to spot (Mahlasela et al., 2024).

Another major issue on social media platforms is the lack of adequate content filtering. Platforms have implemented tools such as content flagging and collaborations with fact-checking groups, but these are sometimes insufficient. It is challenging to adequately control the vast amount of content being posted, and false information keeps spreading. This is demonstrated by the spread of misleading information on COVID-19 and vaccinations throughout the pandemic, which resulted in dangers to the public's health and vaccine reluctance (Demuyakor and Opata, 2022). The spread of social media misinformation emphasises the need for a diversified strategy to resolve the problem. This includes greater algorithm transparency, more effective content filtering, media literacy education, and rules that strike a balance between free expression and the need to protect public discourse (Shu et al., 2020).

4.3 Europe vs Africa Comparative Analysis

Table 1: Comparative Analysis

Category	Europe (Selected Countries)	Africa (Selected Countries	References
Political Systems	Europe has a variety of democratic systems with strong institutions, while certain countries are experiencing rising populism and political polarization (e.g., Italy, Hungary, Czech Republic). The institutional strength varies, with persistent threats from far-right parties and Russian involvement in Eastern Europe.	African political systems are many, ranging from democracies with competitive elections (Ghana, Botswana) to hybrid regimes and authoritarian republics (Cameroon, Uganda). Institutional quality and governance differ significantly, affecting media freedom and misinformation dynamics.	(Jungherr and Schroeder, 2021; Africa Check, 2025; Tandoc, Lim and Ling, 2023)
Media Landscapes	Well-established public broadcasters (such as the BBC in the United Kingdom and Deutsche Welle in Germany) coexist with private media, which are controlled by national government to ensure standards and media integrity. However, media polarization and trust difficulties remain in several countries.	Media environments range from state-controlled outlets (e.g., Ethiopia) to vibrant independent media (e.g., South Africa, Kenya). Informal and digital media play a growing role, especially where traditional media access is limited. Press freedom varies significantly.	(Hameleers and Minihold, 2022; Africa Check, 2025; Tandoc, Lim and Ling, 2023)
Social Media Influence	Fact-checking organizations (e.g., Full Fact in the UK) actively monitor misinformation on platforms like Facebook and X (Twitter). Algorithmic amplification and political polarization contribute to misinformation spread, especially during elections.	Social media is a critical source of information, often filling gaps left by traditional media. WhatsApp and Facebook are widely used for news sharing, but weak regulation and lower digital literacy can accelerate misinformation spread, notably in countries like Nigeria and South Africa.	(Carmi et al., 2020; Africa Check, 2025; Tandoc, Lim and Ling, 2023)

Category	Europe (Selected Countries)	Africa (Selected Countries	References
Digital Literacy	High digital literacy rates are supported by formal education systems emphasizing media literacy. Citizens generally have awareness of misinformation tactics, though gaps remain among older populations	Digital literacy varies significantly; urban centers tend to have higher digital skills, while rural areas often have limited access and education on misinformation. Efforts to improve digital literacy are growing but uneven across countries.	(Carmi et al., 2020; Africa Check, 2025; Tandoc, Lim and Ling, 2023)
Disinformation Challenges	Europe faces sophisticated misinformation campaigns, including Russia-linked operations, deepfakes, and AI-generated fake news, particularly affecting Eastern Europe and during electoral cycles.	Challenges include political propaganda, state-sponsored disinformation, and misinformation spread through encrypted messaging apps. Conflict zones and politically unstable countries (e.g., Cameroon, Ethiopia) are especially vulnerable.	(Jungherr and Schroeder, 2021; Africa Check, 2025; Tandoc, Lim and Ling, 2023; UNESCO, 2025)
Fact-Checking & Regulation	Supported by comprehensive EU regulations such as the Digital Services Act (DSA) and GDPR, alongside active fact-checking networks across member states. Regulatory enforcement varies but overall frameworks are robust.	Regulatory frameworks are emerging but often underdeveloped. Initiatives like Africa Check and localized fact-checking efforts are growing, but enforcement and capacity remain limited in many countries. Regional bodies like the African Union are promoting media governance reforms.	(Hameleers and Minihold, 2022; Africa Check, 2025; Tandoc, Lim and Ling, 2023; UNESCO, 2025)

4.4 Role of Fact-Checking Initiatives in Both Regions

Fact-checking plays a significant role in reducing false beliefs as a way of combating the spread of misinformation. Fact-checkers are tasked with the role of checking whether content produced or posted online is factual. A study by Porter and Wood (2021) focused on assessing the effectiveness of global fact-checking, discusses how fact-checking can help deny false beliefs, but it is not enough to eliminate or reduce misinformation. A study by Badji, Orgeret and Mutsvairo (2024) on fact-checking practices in Ethiopia and Mali discusses how fact-checkers in these countries are often victims of harassment and online bullying, who self-censor and avoid working on sensitive topics due to fear of reprisal from the government. They further state how this leads to inaccessibility of reliable resources and their focus shifting to viral social media content rather than focusing on being truth keepers (Badji, Orgeret and Mutsvairo, 2024). Resource constraints in both regions are a challenge, as fact-checkers do not have access to the resources that will help in eliminating misinformation due to politics. Digital transformation is a challenge where there is a growing number of organizations that are responsible for fact-checking, but there is a lack of media literacy in both regions (García-Gordillo, Rivas-de-Roca and de-Lima-Santos, 2025).

4.5 Role of Deepfakes in Worsening Misinformation

The emergence of AI has made it easier for deepfakes to be more realistic, thus making it harder for people to distinguish between fake and real content (Mahlasela et al., 2024; Veerasamy and Pieterse, 2022). Due to the realistic nature of deepfakes, it has made it hard for people to tell what is right and wrong due to the realistic audio, videos and images. This has resulted in the rapid spread of deepfake content on social media as people do not verify the authenticity of the content (Baloyi et al., 2024; Luleci and Catbas, 2023). There has been the emergence of new tactics of spreading misinformation using people in high positions, politicians, due to the emergence of deepfakes. For instance, there was a case where a South African parliamentary platform was used to spread misinformation on how the parliament is in support of Bitcoin (Panchia, 2025). Deepfakes have made it easy for malicious actors to spread misinformation, as there will be new tactics that are used to spread information through social media platforms such as Twitter and Facebook (Perez Dasilva, Meso Ayerdi and Mendiguren Galdospin, 2021).

According to Gregory (2022) social media has been used as a form of news media to reach the masses. Therefore, deepfakes have been used for media manipulation in spreading misinformation through the manipulation of images, voices, and videos of people in high positions or prominent figures to push a narrative. Veerasamy and Pieterse (2022) how deepfakes use prominent figures such as Donald Trump and Mark Zuckerberg to cause confusion among users or to create an influence. This further leads to distrust amongst users. Luleci and Catbas (2023) discuss how deep-fake technology poses challenges for detection due to the sophisticated nature of deepfakes, resulting in the spread of misinformation, as it is difficult to distinguish between real and manipulated content. False narratives are amplified using deepfakes to trigger emotional responses, create bias, and exploit societal views through manipulation (Luleci and Catbas, 2023).

4.6 Current State of Deepfake Technology and Detection

The discussion of how media manipulation is due to deep fakes is a growing concern due to the rise of Artificial intelligence (Mahlasela et al., 2024). The authors further discuss how creating more diffusion models is used to create realistic content for purposes that were not intended, such as image creation. Generative adversarial networks (GANs), a deep learning model used to create realistic images and videos, make it difficult to detect deepfakes due to their adaptability, which allows them to be used to train any data type (Mahlasela et al., 2024). Deepfake tools' ease of access and usage make it easy to create content, as anyone can have access to open-source tools (Mahlasela et al., 2024; Du et al., 2023). AI has made it easy for the creators of deepfakes to use new techniques and tools that can create realistic images, audio, and videos. AI has made creating deepfakes easy for anyone without experience or skill. Cinar (2023) further emphasizes how deepfakes are used during elections to push an agenda and affect democracy and the rights of citizens due to the creation of misleading images and videos of politicians spreading their agenda.

5. Discussion

The rapid spread of misinformation on social media presents distinct challenges in Europe and Africa that are shaped by the differences in the regulatory frameworks, digital literacy, political landscapes, and technological infrastructure. The spread and impact of misinformation vary significantly between Europe and Africa due to the distinct political, technological, and socio-economic factors. Political polarization and algorithmic amplification in Europe play a major role in driving misinformation. Social media platforms like Facebook and Twitter focus on engagement, which often amplify divisive content that seeks to reinforce ideological echo chambers (Hameleers and Minihold, 2022). While the European Union has implemented regulatory measures like the Digital Service Act (DSA) and the Code of Practice on Disinformation to increase platform transparency to ensure enforcement remains inconsistent (European Commission, 2022). Also, higher levels of digital literacy in Europe help mitigate misinformation, though ideological biases still influence media consumption (Van Kessel, Sajuria and Van Hauwaert, 2020).

On the other hand, Africa faces different challenges primarily due to reliance on informal networks and weak regulatory frameworks. Messaging platforms like Facebook and WhatsApp dominate with limited content moderation, allowing rumors and false information to spread unchecked (Mtsweni et al., 2023). In politically unstable democracies such as Nigeria and Kenya, misinformation is frequently weaponized during the elections, which often incites violence (Moyo, Mare and Mabweazara, 2020). Fact-checking initiatives like Africa Check struggle with insufficient funding, while governments sometimes resort to extreme measures such as internet shutdowns to curb misinformation (Yussif et al., 2024). These disparities highlight how regional differences in technology, governance and media literacy shape the spread of and control of misinformation.

The emergence of deepfakes has led to manipulation, confusion and mistrust in organizations by users. A study by (Mahlasela et al., 2024) focused on the impact of artificial intelligence on the realism and prevalence of deepfakes. As emphasized in this study, GANs, FBMs and ARs are used to create realistic images and videos, which makes it difficult to detect deepfakes (Mahlasela et al., 2024). The advancement in AI has made it easy for threat actors to create undetectable deepfakes. Cinar (2023) in their study emphasized how deepfakes are used to push an agenda, such as during elections, where videos and images of politics are used to create a narrative. This poses a threat as these deepfakes are used to spread misinformation.

Due to the increasing use of social media, threat actors have found new ways of targeting their victims. Deepfakes help with the spreading of misinformation due to the sophisticated nature of deepfakes (Al-Khazraji et al., 2023). This leads to mistrust and confusion amongst users as they are unable to verify real and fake content. Open-access tools have made it easy for anyone to create deep fakes. This leads to an increase and improvement of deepfakes, which makes it hard to detect what's real and fake. As emphasized in the studies by Baloyi et al., 2024; Al-Khazraji et al. 2023), deepfakes have made it difficult for users to detect what's real and fake, thereby sharing unverified content online on social media. This further leads to bias, emotional triggers and exploitation of societal views through manipulation (Al-Khazraji et al., 2023).

6. Misinformation Framework

Social media has made it easy to spread misinformation, and with the popularity of deepfakes, misinformation has worsened (Al-Khazraji et al., 2023). This section discusses the proposed misinformation framework as illustrated in Figure 2 on how to combat misinformation and educate society on how not to fall victim to misinformation.



Figure 2: Proposed Misinformation Framework

6.1 Multi-Stakeholder Partnership

Misinformation calls for a multi-stakeholder partnership with different stakeholders. This includes governments enforcing laws and regulations that will protect the citizens and promoting media literacy on identifying misinformation. Private sector organisations, mainly focusing on technology, can develop detection tools that will be used to detect deep-fakes, which will create a media-literate society. These will help with transparency purposes when accessing social media platforms. Media organisations will play a massive role in engaging in more fact-checking efforts and accurate reporting by discarding fake social media platforms aimed at disseminating misinformation and ensuring adherence to ethical standards. In addition, educational institutions could include a curriculum on media literacy. Research institutions can conduct more research on misinformation and develop new detection techniques while collaborating with media organisations to share educational content through the media platforms. This will create a more digitally literate society where people know how to identify misinformation.

6.2 Cross-regional Collaboration

Collaboration has been seen as an essential approach when faced with challenges. This will include sharing information between the two regions, such as making information open access so that people in different regions have access to it. A study conducted by Badji, Orgeret, and Mutsvairo (2024) shared how fact-checking organisations do not have access to resources that can be used to eliminate misinformation in politics. Sharing detection techniques in both regions will help tackle misinformation, with instances where Africa will learn from European organisations on deepfake detection techniques. Collaboration in research aimed at experts in the regions, coming together to study misinformation, will tackle misinformation. This can also include different experts who specialise in misinformation and develop solutions that will tackle misinformation. Fostering global partnerships between educational institutions, research institutions, international organizations, and governments to combat misinformation globally will help create an informed global society. International organisations should collaborate to develop international standards and laws that are aimed at addressing misinformation.

6.3 Deepfake Detection Tools

Governments in both regions should invest in research efforts to develop advanced deepfake detection tools. This will further promote collaboration between Europe and Africa, where different stakeholders work together to develop and implement deepfake detection tools. Private sector organisations can partner with public sector organisations to create deep-fake detection tools that can be integrated into social media platforms. These detection tools should abide by the laws and regulations that will ensure the effective implementation of deepfake detection tools.

6.4 More Informed Digital Society

Globally, countries should emphasise the importance of taking media literacy programs starting from primary school. These programs will educate the public on how to identify misinformation and disinformation, follow the right social media platforms that disseminate information, detect deepfakes, and evaluate information sources, including fact-checking efforts. This will empower citizens to participate in misinformation elimination efforts by creating misinformation advocates who are tasked with raising awareness on how to identify misinformation. This can include providing free access to tools that can verify information and encouraging fact-

checking before sharing information online. Emphasis on the consequences of sharing information without verification can encourage citizens to create awareness of misinformation, especially during elections.

7. Conclusion and Future work

This study analysed the key factors influencing the spread of misinformation on social media platforms in Europe and Africa. It further investigated the role of social media, users, bots, and other digital tools in amplifying misinformation. In addition, it aimed to identify similarities and differences in misinformation trends, propagation mechanisms, and audience susceptibility in Europe and Africa. The results indicated that it is essential to implement the proposed framework to combat misinformation and eliminate the continuous spread of misinformation on social media. For future work, researchers can further explore how misinformation affects Africa by comparing specific countries in Africa that are developing and developed.

Ethical declaration: No ethical clearance was required as this study involved a systematic literature review.

AI declaration: No AI tools were used in the development of this research paper.

References

Access Now. (2020) *Nigeria withdraws controversial social media bill — victory for free expression*. Available at: https://www.accessnow.org/press-release/nigeria-withdraws-controversial-social-media-bill-victory-free-expression/ (Accessed: 28 March 2025).

Africa Check. (2023) *Fact-checking in Africa: progress and challenges*, Johannesburg: Africa Check. Available at: https://africacheck.org/fact-checking-in-africa (Accessed: 14 July 2025).

Ahmed, S., Madrid-Morales, D. and Tully, M. (2023) "Social media, misinformation, and age inequality in online political engagement", *Journal of Information Technology & Politics*, Vol 20, No. 3, pp 269–285.

Aïmeur, E., Amri, S. and Brassard, G. (2023) "Fake news, disinformation and misinformation in social media: a review", *Social Network Analysis and Mining*, Vol 13, No. 1, pp 1–36.

Al-Khazraji, S. H., Saleh, H. H., Khalid, A. I. and Mishkhal, I. A. (2023) "Impact of deepfake technology on social media: Detection, misinformation and societal implications", *The Eurasia Proceedings of Science Technology, Engineering and Mathematics*, Vol 23, pp 429-441.

Badji, S. D., Orgeret, K. S. and Mutsvairo, B. (2024) "An Exploratory Study of Fact-Checking Practices in Conflict and Authoritarian Contexts", *Media and Communication*, Vol 12.

Baqir, M., Kennedy, M. and Albright, C. (2024) "Cross-Country Comparison of News and Misinformation Consumption in Europe: A Longitudinal Study", *Journal of Communication Studies*, Vol 23, No. 4, pp 89-103.

Baloyi, E., Mahlasela, O., Siphambili, N. and Stegmann, M. (2024) "Social Media as a Strategic Advantage during Cyberwarfare: A Systematic Literature Review", *International Conference on Cyber Warfare and Security*, Vol 19, No. 1, pp 19–25.

BBC News. *Trending*, Available at: https://www.bbc.com/news/blogs-trending-38510353 (Accessed: 20 March 2025).

Bhardwaj, A. and Kumar, V. (2023) "Social media as an enabler to combat misinformation", in *Pandemics in the Age of Social Media*, Routledge, pp 1-15.

Bouchaud, P. (2023) "Algorithmic amplification of politics and engagement maximization on social media", in *International Conference on Complex Networks and Their Applications*, Cham: Springer Nature Switzerland, pp 131-142.

Carmi, E., Yates, S. J., Lockley, E. and Pawluczuk, A. (2020) "Data citizenship: Rethinking data literacy in the age of disinformation, misinformation, and malinformation", *Internet policy review*, Vol 9, No. 2, pp 1-22.

Cazzamatta, M. (2024) "Global Misinformation Trends: Comparative Analysis of Europe and Africa", *Journal of Media & Politics*, Vol 12, No. 2, pp 91-108.

Cazzamatta, R. (2024) "Global misinformation trends: Commonalities and differences in topics, sources of falsehoods, and deception strategies across eight countries", *New Media and Society*. Available at: https://doi.org/10.1177/14614448241268896/SUPPL_FILE/SJ-DOCX-1-NMS-10.1177_14614448241268896.DOCX.

Chenzi, V. (2021) "Fake news, social media and xenophobia in South Africa", *African Identities*, Vol 19, No. 4, pp 502–521.

Cinar, B. (2023) "Deepfakes in Cyber Warfare: Threats, detection, techniques and countermeasures", *Asian Journal of Research in Computer Science*, Vol 16, No. 4, pp 178-193.

De Jager, A., Marivate, V. and Modupe, A. (2023) "Multimodal Misinformation Detection in a South African Social Media Environment", in *Southern African Conference for Artificial Intelligence Research*, Cham: Springer Nature Switzerland, pp 285-299.

Demuyakor, J. and Opata, E. M. (2022) "Fake news on social media: predicting which media format influences fake news most on facebook", *Journal of Intelligent Communication*, Vol 1, No. 2.

Di Domenico, G., Sit, J., Ishizaka, A. and Nunan, D. (2021) "Fake news, social media and marketing: A systematic review", *Journal of Business Research*, Vol 124, pp 329–341.

Di Nubila, K., Ballesteros-Herencia, C. A., Etura, D. and Martín-Jiménez, V. (2023) "Technopopulism and Politainment in Brazil: Bolsonaro Government's Weekly YouTube Broadcasts", Media and Communication, Vol 11, No. 2, pp 137-147.

Du, H., Niyato, D., Kang, J., Xiong, Z., Zhang, P., Cui, S., Shen, X., Mao, S., Han, Z., Ja-malipour, A. and Poor, H.V. (2023) The age of generative AI and AI-generated everything, arXiv preprint arXiv:2311.00947.

European Commission. (2022) Digital Services Act: strengthening the responsibility of online platforms. Available at: https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package (Accessed: 14 July 2025). European Commission. (2022) Digital Services Act. Available at: https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en (Accessed: 31 March 2025).

Film and Publication Board. (2024) Notice. Available at: https://www.gov.za/sites/default/files/gcis_document/202403/50311gon4526.pdf (Accessed: 29 March 2025).

Fraga-Lamas, P. and Fernandez-Carames, T. M. (2020) "Fake news, disinformation, and deepfakes: Leveraging distributed ledger technologies and blockchain to combat digital deception and counterfeit reality", IT professional, Vol 22, No. 2, pp 53-59.

García-Gordillo, M., Rivas-de-Roca, R. and de-Lima-Santos, M. F. (2025) "The fact-checking initiatives in the EU: a diverse ecosystem against disinformation", Media and Communication, Vol 13.

Gondwe, C., Kamanga, D. and Mbekeani, M. (2025) "Misinformation and Digital Inequalities in Africa: A Cross-National Study", Journal of African Media Studies, Vol 14, No. 2, pp 153-175.

Gregory, S. (2022) "Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism", Journalism, Vol 23, No. 3, pp 708-729.

Hameleers, M., Brosius, A. and de Vreese, C. H. (2022) "Whom to trust? Media exposure patterns of citizens with perceptions of misinformation and disinformation related to the news media", European journal of communication, Vol 37, No. 3, pp 237-268.

Hameleers, M. and Minihold, J. (2022) "Disinformation, Polarization, and Right-Wing Populism in Europe: A Comparative Perspective", European Journal of Political Science.

Jungherr, A. and Schroeder, R. (2021) "Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy", Social Media+ Society, Vol 7, No. 1, 2056305121988928.

Kestler, T. (2022) "Radical, nativist, authoritarian—or all of these? Assessing recent cases of right-wing populism in Latin America", Journal of Politics in Latin America, Vol 14, No. 3, pp 289-310.

Luleci, F. and Catbas, F.N (2023) "A brief introductory review to deep generative models for civil structural health monitoring", AI Civ. Eng., Vol 2, 9.

Mahlasela, O., Baloyi, E., Siphambili, N. and Zubeida, K. (2024) Artificial Intelligence Impact on the realism and prevalence of deepfakes. Available at: https://www.researchgate.net/publication/384056978.

Mare, A., Mabweazara, H. M. and Moyo, D. (2019) ""Fake News" and Cyber-Propaganda in Sub-Saharan Africa: Recentering the Research Agenda", African Journalism Studies, Vol 40, No. 4, pp 1–12.

Morosoli, S., Van Aelst, P., Humprecht, E., Staender, A. and Esser, F. (2025) "Identifying the drivers behind the dissemination of online misinformation: A study on political attitudes and individual characteristics in the context of engaging with misinformation on social media", American Behavioral Scientist, Vol 69, No. 2, pp 148-167.

Moyo, D., Mare, A. and Mabweazara, H.M. (2020) "Social media, the press, and the crisis of disinformation in Africa", Communication: South African Journal of Communication Theory and Research, Vol 46, No. 4, pp 1-6.

Mtsweni, J., Ndlovu, L., Mthethwa, S. and Mkuzangwe, N. (2023) "Measuring misinformation trends on social media in South Africa using Machine Learning", in International Conference on Artificial Intelligence and its Applications, pp 128-133.

Onigbinde, C. O. and Oloyede, I. B. (2024) "Misinformation and disinformation in the Nigerian media landscape", IMSU Journal of Communication Studies, Vol 8, No. 1, pp 93-100.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A. and Brennan, S. E. (2021) "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews", BMJ, Vol 372.

Panchia, Y. (2025) SA Parliament's social media accounts hacked to promote the launch of 'Ramaphosa token', Daily Maverick, 15 March. Available at: https://www.dailymaverick.co.za/article/2025-03-15-sa-parliaments-social-media-accounts-hacked-to-promote-the-launch-of-ramaphosa-token/ (Accessed: 29 March 2025).

Perez Dasilva, J., Meso Ayerdi, K. and Mendiguren Galdospin, T. (2021) "Deepfakes on Twitter: which actors control their spread?", Media and Communication, Vol 9, No. 1, pp 301-312.

Porter, E. and Wood, T. J. (2021) "The global effectiveness of fact-checking: Evidence from simultaneous experiments in Argentina, Nigeria, South Africa, and the United Kingdom", Proceedings of the National Academy of Sciences, Vol 118, No. 37, e2104235118.

Shu, K., Mahudeswaran, D., Wang, S. and Liu, H. (2020) "Hierarchical propagation networks for fake news detection: Investigation and exploitation", in Proceedings of the international AAAI conference on web and social media, Vol 14, pp 626-637.

Tandoc, E.C. Jr., Lim, Z.W. and Ling, R. (2023) 'Defining "fake news": a typology of scholarly definitions', Digital Journalism, Vol 11, No. 3, pp 345-363. doi: 10.1080/21670811.2022.2126667.

The Guardian. 85% of people worry about online disinformation, global survey finds. Available at: https://www.theguardian.com/technology/2023/nov/07/85-of-people-worry-about-online-disinformation-global-survey-finds (Accessed: 01 April 2025).

UNESCO. (2022) Global media and information literacy assessment framework, Paris: UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000373453 (Accessed: 14 July 2025).

Van Kessel, S., Sajuria, J. and Van Hauwaert, S. M. (2020) "Informed, uninformed or misinformed? A cross-national analysis of populist party supporters across European democracies", West European Politics, Vol 44, No. 3, pp 585-610.

Van Tilburg, L. (2025) South African soldiers are essentially being held as POWs in the DRC - National Security News, National Security News, 4 February. Available at: https://nationalsecuritynews.com/2025/02/sa-soldiers-in-drc-practically-pows-claims-leading-analyst/ (Accessed: 27 March 2025).

Veerasamy, N. and Pieterse, H. (2022) "Rising above misinformation and deepfakes", in International Conference on Cyber Warfare and Security, Vol 17, No. 1, Academic Conferences International Limited, pp 340-348.

Vikranth, K., Nethravathi, P.S. and Krishna Prasad, K. (2024) "Price prediction system – a predictive data analytics using arima model", Ictact Journal on soft computing, Vol 14, No. 3, pp 3304-3310.

Wasserman, H. and Madrid-Morales, D. (2019) "An Exploratory Study of "Fake News" and Media Trust in Kenya, Nigeria and South Africa", African Journalism Studies, Vol 40, No. 1, pp 107–123.

Wu, L., Morstatter, F., Carley, K.M. and Liu, H. (2019) "Misinformation in social media: definition, manipulation, and detection", ACM SIGKDD explorations n

Yussif, A.B., Agyepong, L., Odoom, D., Tindi, S. and Dick-Sagoe, C. (2024) "Social Media and Political Communication in Ghana: Views of Inhabitants of Tema Metropolis", Information Impact: Journal of Information and Knowledge Management, Vol 15, pp 30–46.

Revisiting Biometrics in Cybersecurity: Do AI Methods and Zero-Trust Architectures Drive Innovation?

Siphesihle Sithungu[1] and Christoph Lipps[2]
[1]Academy of Computer Science and Software Engineering, University of Johannesburg, South Africa
[2]Intelligent Networks Research Group, German Research Center for Artificial Intelligence, Kaiserslautern, Germany

siphesihles@uj.ac.za
christoph.Lipps@dfki.de

Abstract: Biometric authentication has long been regarded as a foundational element of identity verification, leveraging unique physiological and behavioral traits to enhance security beyond traditional passwords. While it offers notable advantages such as convenience and resistance to identity theft, concerns are mounting regarding privacy, susceptibility to spoofing, and the irreversibility of compromised biometric identifiers. These weaknesses are becoming increasingly critical as digital infrastructures evolve into distributed, dynamic environments in which static trust models are no longer sufficient. Moreover, several traditional modalities -such as fingerprints, iris scans, and voice recognition- have already been breached. However, Artificial Intelligence (AI) methods are reshaping this landscape by introducing adaptive and context-aware features into biometric systems. Machine Learning (ML) techniques enhance accuracy, enable continuous authentication, and support multimodal fusion, while anomaly-detection mechanisms improve resilience against sophisticated attacks. Generative AI (GenAI) plays a particularly significant role, though it introduces a paradox: it empowers defenders through realistic attack simulations and robustness testing, yet simultaneously equips attackers with tools for producing deepfakes and synthetic identities, thereby expanding the attack surface. In this evolving security landscape, Zero-Trust Architectures (ZTA) have gained prominence as a model that replaces assumptions of inherent trust with continuous verification mechanisms. The use of biometric data within ZTA can enhance the reliability of identity verification; however, it also intensifies several existing issues. Biometric identifiers must be handled and stored in ways that safeguard individual privacy and align with relevant legal requirements, and the incorporation of AI-based assessment methods introduces additional concerns regarding potential bias, transparency, and oversight. Moreover, combining AI-supported biometric systems with Zero-Trust principles raises further questions about scalability, system compatibility, and the broader ethical consequences of more pervasive identity monitoring. This work therefore examines the convergence of biometrics, AI, and Zero-Trust principles from a critical perspective. It highlights the dual role of AI as both a source of innovation and a generator of new threats, while identifying opportunities for adaptive security, real-time threat detection, and improved user experience. By analyzing technical and operational dimensions, the work proposes a roadmap for integrating biometrics into ZTA that balances innovation with accountability and supports trustworthy, resilient cybersecurity frameworks.

Keywords: Biometrics, Zero-trust architecture, Generative AI, Machine learning, Cybersecrutiy

1. Biometric Authentication: From Face Recognition to AI-driven Identity Verification

Current developments in Artificial Intelligence (AI) methods are -again- shifting biometric authentication from the fringes of research to everyday security. Facial recognition unlocks smartphones and payments, fingerprint readers secure workplaces, and voice biometrics speed up customer support. With applications increasingly distributed across clouds and devices, identity has become the new security perimeter, and biometrics provide a human-centric authentication factor improving both usability and security.

This development did not happen overnight, but rather biometric technologies have gone through four phases of development: (i) early research in the 1960s; (ii) large-scale government implementations in the 1990s and 2000s; (iii) consumer adoption via smartphones in the 2010s; and (iv) today's AI-driven boom in the 2020s (Jain, Nandakumar, and Ross, 2016)(Buriro, & Luccio, 2025)(Wayman et al. 2025). Thereby, each phase was triggered by advances and changing security requirements shaping this landscape.

And now, the further developments of AI methods have again changed both, the accuracy and speed of biometric recognition, enabling systems to operate at scale and adapt to different conditions. But these same advances have also opened new attack vectors: deepfakes, high-resolution voice clones, adversarial inputs, and synthetic identities (Farooq et al., 2025). This duality raises a central question: *Do current AI-based methods offer security innovations, or do they merely replace old weaknesses with new ones?*

Historically, biometric systems were designed for closed, static environments with cooperative users and calibrated sensors, as well as risk models balancing false acceptance and false rejection. But, today, authentication takes place in dynamic, distributed contexts: Bring-your-own-device (BYOD) endpoints, variable networks and sensors, evolving user behavior, and continuous software updates. Attackers exploit these

conditions with automated, low-cost tools. In such environments, static thresholds and point-in-time checks are insufficient. What matters is continuous risk assessment, rapid model adaptation, origin-aware detection, and deep defense that assumes compromise, a mindset embodied by Zero Trust. AI methods, on the other hand, are driving this change forward for both defenders and attackers.

On the defense side, deep learning (DL) has improved detection accuracy, enhanced liveness detection, enabled cross-modality fusion, and supported continuous authentication through behavioral biometrics such as keystroke dynamics and gait. On the attacker side, generative models reduce the cost and expertise required to create convincing fakes and adversarial examples. Added to this is the large-scale sharing of biometric data -from photos to voice samples- which highlights a biometric problem: irreversibility. Unlike passwords, biometric features cannot be rotated after a compromise; they can be converted into erasable templates or bound to additional secrets, but the original signal remains visible or is lost forever (Lipps, Herbst, and Schotten, 2021).

This reveals a major gap between research and practice: Many implementations are still based on static statistical models with questionable assumptions about distributions and user cooperation. Calibration drifts as populations diversify and contexts evolve; thresholds tuned into laboratories deteriorate in production; liveness checks that prevent simple spoofing fail against AI-assisted attacks. Furthermore, companies often use biometrics as their sole access control rather than as a risk signal within a broader zero-trust framework, where identity assurance is contextual, continuously updated, and combined with device integrity, network configuration, and behavioral analytics. Thereby, the result is a misalignment as AI-powered biometrics is marketed as an end-to-end solution in environments requiring composable, continuously verifiable signals.

The critical question therefore is not whether AI increases accuracy -it does- but whether it increases security assurance against adaptive attackers and real-world constraints. Any evaluation must go beyond closed benchmarks and focus on operational robustness: resilience to presentation and injection attacks, transferability of adversarial examples, resilience to distribution shifts, privacy-compliant template management, and graceful degradation in the event of partial compromise. In parallel, biometrics must be transformed from immutable identifiers to revocable, policy-bound signals within zero-trust architectures: anchored in device attestations, combined with cryptographic protocols, and continuously monitored with anomaly detection and explainable risk assessment (Uppal, et al., 2024).

Against this backdrop, this work essentially pursues three goals: First, to highlight the research gap between laboratory results and operational requirements for resilience, recoverability, and trustworthy calibration in dynamic, distributed environments. Second, to analyze the dual role of AI -how it enables stronger detection and liveness defense while introducing new classes of attacks and risks in the supply chain. and third, to position biometrics within Zero Trust as a composable signal rather than a single point of truth, outlining architectural patterns that integrate biometrics with device trust, continuous authentication, and privacy-preserving design. Ultimately, innovation in AI-driven biometrics is less about incremental gains in recognition benchmarks and more about end-to-end security engineering: rigorous threat modeling, measurable robustness, revocability in the event of compromise, built-in privacy, and principled integration into Zero Trust workflows. Redefining biometrics from this perspective separates hype from progress and sets a research agenda that strengthens digital trust, not just rationalizes it.

2. The Background: From Traditional Biometrics to Zero-Trust Principles

Biometric authentication has developed from the use of physical traits such as fingerprints and iris patterns to advanced behavioural and continuous authentication methods. Its evolution reflects increasing computational capabilities, usability demands and the shift toward probabilistic, ongoing identity verification within Zero Trust frameworks.

2.1 Definition of Biometric Authentication

Biometric authentication is a process making use of unique physiological or behavioural characteristics of entities (i.e. fingerprints, voice patterns, facial features, gait, iris scans) to determine the legitimacy of access they claim to have in a particular system (Amiri, et al., 2024). In computing, biometrics are considered a more reliable access control mechanism as opposed to passwords (which rely on memory) or tokens (which rely on possession) as authentication is linked directly to the individual (Ryu, et al., 2023). Figure 1 illustrates the three levels into which authentication factors are divided.

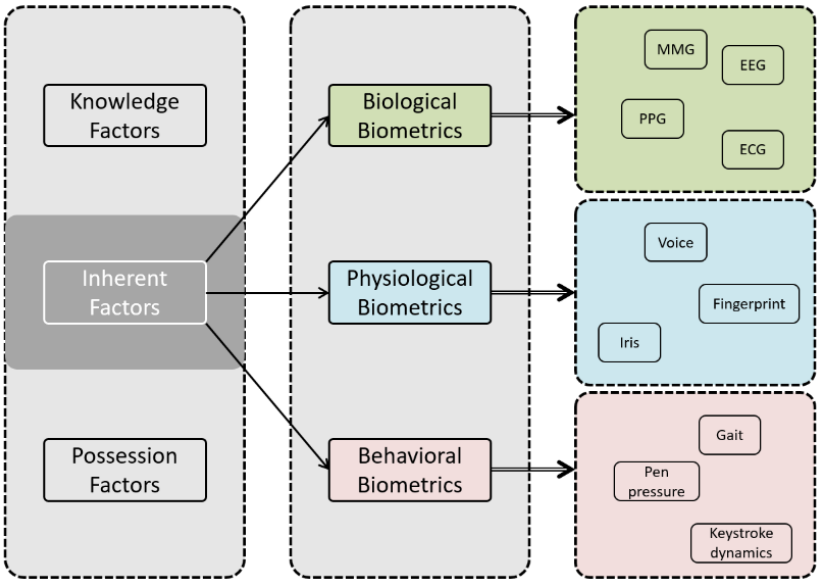


Figure 1: Authentication Factors are basically divided into three levels: Knowledge-, Inherence- and Possession Factors. Inherence Factors are further specified into the sub-groups: biological-, physiological- and behavioural biometrics, which could be further specified (Lipps, et al., 2022)

As indicated in Figure 1, authentication traditionally relies on three categories of factors: knowledge-based, possession-based, and inherence-based factors. Knowledge-based factors depend on information the user knows, such as passwords or PINs, whereas possession-based factors require something the user physically holds, such as a smart card or security token. Inherence-based factors, by contrast, rely on attributes inherent to the individual. Biometrics therefore fall within this third category and can be further differentiated, for instance, into biological, physiological, and behavioural characteristics. Examples of these inherent traits -such as fingerprints, iris patterns, voice characteristics, or gait- are illustrated in the right-most part of the Figure (Lipps, et al., 2022).

Due to advances made in the field of computer science within the last two decades, biometric authentication systems may also apply learning mechanisms to improve pattern recognition. The ability to learn also enables biometric systems to adapt over time and refine their ability to detect biometric features (Amiri, et al., 2024).

2.2 Evolution of Biometrics from Fingerprints and Iris Scans and the Rise of AI-based Methods

Early biometric systems progressed from measuring what a person *touches*, to what can be *seen*, and eventually to how a person *behaves*. This evolution was primarily driven by the desire to capture richer information, reduce user friction, and leverage increasing computational capabilities. The earliest widely adopted biometric modality was fingerprint recognition, due to its high distinctiveness, long-term stability, and the relative simplicity of comparison algorithms. However, fingerprint-based systems also faced notable limitations as they required direct physical contact, raised hygiene concerns, and were vulnerable to physical wear or intentional alteration of the skin. These challenges were less about accuracy and more about usability and practicality in real-world environments. (Herbst et al., 2024).

The next wave of biometrics focused on iris and retina scans, which moved the medium to “inside” the body where information is richer and replication is more difficult, resulting in a significant improvement in accuracy. However, the biggest trade-off was cost (i.e. specialised sensors and controlled lighting) and intrusiveness (user discomfort). These factors made iris detection suitable for high-security environments (i.e. border control) and less practical for most use cases. Furthermore, face recognition became the most ideal option in many cases due to the increasing availability of high-quality cameras and the advantages that come from using faces. Faces are socially natural, do not require physical contact and inexpensive to capture. The associated costs (lighting, ageing, pose and bias) resulting in this kind of detection being statistical rather than absolute.

More recently, biometrics evolved into voice and behaviour where voice biometrics analyse pitch, cadence and spectral features whereas behavioural biometrics analyse a range of behaviours depending on context (i.e. typing or swiping style, gait or how you hold a mobile device). Such biometrics are ongoing and passive, turning authentication from a one-off activity into an ongoing process. The idea of continuous authentication has also

led to another emerging security paradigm (canonically unrelated to biometrics) called Zero Trust, which is described in the following section.

2.3 What is Zero Trust

Zero Trust -often referred to as Zero Trust Architecture (ZTA)- emerged within systems security and is fundamentally concerned with answering a key question: "Given a stream of requests, what is the likelihood that this is still the same user?" Because this question is both probabilistic and time-dependent, it cannot be resolved through a single authentication event conducted prior to granting access. Instead, it necessitates a model of continuous authentication.

Traditional perimeter-based security operated under the assumption of a clear boundary between the "inside" and the "outside." This model collapsed with the rise of cloud computing, mobile devices, and the increasing prevalence of compromised credentials. A one-time login check is no longer adequate to establish -or maintain- trust. As outlined in guidance from the National Institute of Standards and Technology (NIST), Zero Trust shifts the focus toward safeguarding resources and continuously evaluating trust rather than granting it implicitly (Rose, et al., 2020).

A notable connection between ZTA and behavioural biometrics is that, despite their different origins -behavioural biometrics in signal processing and pattern recognition, and Zero Trust in systems security- both disciplines adopt the principle of continuous authentication. Behavioural biometrics poses a parallel question: "Given a stream of actions, what is the likelihood that this is still the same user?" This, too, is a probabilistic and dynamic assessment. Consequently, both approaches converge on the idea that trust is not binary, confidence naturally degrades over time, evidence arrives continuously, and authentication decisions must remain open to revision.

3. Rethinking Biometric Security in the AI era: Fundamental Risks and the Limits of Traditional Trust Models

Despite significant progress in AI-driven biometrics, critical gaps remain between laboratory performance and real-world security requirements. First, current systems face new vulnerabilities introduced by AI itself, including adversarial attacks, deepfakes, and synthetic identities that undermine trust in biometric authentication. Second, the inherent irreversibility of biometric features means that once compromised, they cannot be easily revoked or replaced, creating long-term exposure risks. Third, many deployments still rely on outdated, static trust models that fail in dynamic, distributed environments where continuous verification and adaptive risk assessment are essential. Addressing these challenges is key to ensuring that AI-driven biometrics deliver genuine security innovation rather than shifting old weaknesses into new forms.

3.1 The Main Issues Considering AI in Biometric Authentication

One key aspect, however, is the dual purpose of today's AI in biometric authentication, leading to a significant conflict between improved accuracy and increased vulnerability to attacks. Generative and discriminative techniques designed to improve recognition performance also lower the barriers to presentation and injection attacks -including deepfakes, voice clones, and synthetic fingerprints-, enabling spoofing to scale and identity theft in remote onboarding and electronic Know Your Customer (eKYC) contexts (Verma et al., 2023). The limitations of current mechanisms for detecting liveness and presentation attacks (PAD) against model-aware, AI-powered attackers point to a shift from artifact-based identity spoofing to hostile manipulation of decision boundaries. At the same time, large-scale aggregation and scraping of biometric data poses significant privacy risks, as compromised features are irreversible and common methods for protecting templates are insufficient to ensure revocation semantics comparable to passwords. Taken together, these factors reveal a security gap between benchmark accuracy and operational robustness, necessitating a transition from accuracy-oriented reporting to risk- and attack surface-oriented design, assessment, and governance.

3.2 The Irreversibility of Compromised Biometric Features

Due to these challenges, the irreversibility of biometric features exacerbates systemic risks in an AI-driven threat landscape. Unlike passwords, physiological features cannot be replaced once disclosed, meaning that any compromise becomes a permanent vulnerability rather than a remediable incident. Model inversion and reconstruction attacks can regenerate facial images, fingerprints, or voice patterns from stolen embeddings, and modern AI techniques further exacerbate this problem, allowing attackers to replicate a user's biometric identity at will (Kim et al., 2024). Without meaningful revocation semantics or mature lifecycle controls, compromised

templates remain exploitable indefinitely in systems based on the same modality. At the same time, advances in generative models enable increasingly precise biometric replay and injection attacks that completely bypass sensors and target authentication pipelines directly. These developments underscore the need to reevaluate biometric data as static anchors of trust and instead integrate them into adaptive, risk-oriented authentication frameworks capable of mitigating the lasting effects of compromised features.

3.3 Outdated, Static Trust Models in Dynamic, Distributed Environments

Nonetheless, conventional biometric authentication frameworks continue to rely on centralized trust models no longer aligned with current decentralized, multi-party identity ecosystems. As authentication processes increasingly involve cloud services, third-party providers, federated identity platforms, and remote onboarding channels, static trust anchors cannot capture the fluidity and interdependence of modern digital infrastructures. At the same time, these systems lack continuous verification and contextual risk assessment, granting long-lasting trust based on a single biometric match, an assumption that is incompatible with session hijacking, device switching, and adaptive threats. The problem is exacerbated by AI-driven attackers who are able to iteratively refine spoofing inputs in response to system feedback, rendering static threat models obsolete. Furthermore, older biometric trust anchors conflict with the fundamental principles of Zero Trust, which requires continuous verification, context awareness, and the minimization of implicit trust. These limitations underscore the need to move beyond static notions of identity and transition to architectures that position biometrics as dynamic signals within a continuous, risk-based Zero Trust framework.

4. Transformation by Artificial Intelligence Methods

The rapid maturation of artificial intelligence (AI) has materially advanced the design, operation, and assurance of biometric systems. In particular, AI makes a continuous, risk-based approach -fully aligned with Zero-Trust principles- both feasible and practical at enterprise scale. Rather than treating authentication as a one-off verdict, AI enables adaptive, context-aware decision-making that evolves with each interaction and adjusts to changing user states and environmental conditions. Concretely, machine learning improves recognition accuracy and robustness; multimodal fusion and anomaly detection increase resilience and fault tolerance; and generative AI (GenAI) augments training, simulation, and stress-testing while also creating new forms of attack that must be mitigated. Operationally, this AI-driven transformation changes how signals are captured, features are represented, uncertainty is handled, and policy decisions are enforced. Signals are no longer evaluated in isolation but as part of a streamed evidence pipeline in which risk scores are updated continuously and policy engines (e.g., ZTA policy decision points) can step up, maintain, or revoke access dynamically. Data governance and privacy safeguards must keep pace with these capabilities to ensure lawful, fair, and explainable use of AI in identity systems.

4.1 Adaptive and Context-Aware Biometric Systems via AI Methods

AI allows biometric systems to progress from static templates to incrementally updated models that adapt to users and contexts over time. Instead of verifying identity only once at login, the system can re-evaluate whether access remains appropriate at any given moment and under specific conditions. This is achieved by continuously incorporating contextual signals -such as device posture and health, geolocation patterns, time-of-day regularities, network characteristics, session history, and recent behavioural markers- into a dynamic risk score. In practice, a user who typically authenticates from a specific laptop and office network may be granted low-friction continuation, whereas a session that suddenly shifts device, IP range, and typing cadence may trigger step-up controls (e.g., re-capture of a face or voice sample, or a possession check).

Architecturally, adaptive systems rely on feature drift monitoring and model-update strategies (e.g., constrained online learning or periodic batch refresh) to avoid overfitting to transient anomalies while still accommodating natural change (e.g., ageing, stress). They also benefit from calibrated uncertainty estimates (e.g., temperature scaling for score calibration) so that downstream policy engines can interpret scores consistently. From an implementation standpoint, policy-as-code mechanisms tie these evolving risk estimates to clear actions: maintain, limit, re-authenticate, or terminate.

These architectural properties translate directly into practical deployments across high-variability settings. In remote learning and assessment, for instance, conditions change frequently; contextual indicators -such as device swaps, location shifts, and irregular activity patterns- become integral to maintaining trust over time (e.g., proctoring augmented with continuous, context-aware checks). In large, distributed enterprises, multi-agent arrangements coordinate sensors, models, and policy components across edge and cloud, allowing risk to be reassessed as sessions move between networks or devices (Shukur, 2017). Across both scenarios, trust in the

biometric signal is situational rather than absolute, adapting to surrounding factors to increase security while minimising unnecessary user friction.

4.2 Machine Learning for Accuracy and Continuous Authentication

Machine learning (ML) enhances biometric systems by enabling them to learn continuously rather than only at the enrolment stage. Models adjust representations and decision boundaries as user characteristics and operating conditions evolve, accounting for factors such as ageing, fatigue, stress, medication effects, ergonomic changes, and shifts in behaviour. At the same time, ML-based systems filter noise and extract more stable, higher-level features from real-world data (Bengio, Courville, & Vincent, 2013). In practical deployments, this means that the system can refine user-specific embeddings over time and calibrate individual thresholds so that specificity remains high without sacrificing sensitivity across diverse populations.

A key consequence of these capabilities is that ML naturally supports continuous authentication. Rather than relying on a single decisive event, the system can update confidence scores at interaction time -whether through keystroke windows, cursor-movement segments, micro-gesture batches, or brief voice snippets. This finer temporal granularity allows early detection of deviations from a user's behavioural patterns and makes it possible to combine low-friction, passive checks with active verifications (such as short face recapture) only when the assessed risk increases. The result is a model of risk-proportional friction, which enhances security while minimising unnecessary user interruption.

To operate reliably in this dynamic setting, ML-driven biometric systems must incorporate several design safeguards. They require mechanisms to prevent catastrophic forgetting, ensuring that new behavioural observations do not overwrite stable historical patterns. They also need concept-drift detection to distinguish natural behavioural evolution from adversarial manipulation or account compromise. In addition, systems must integrate fairness and bias monitoring to maintain equitable performance across demographic groups, as well as explainability artefacts -such as feature-attribution summaries or confidence decompositions- that support governance, auditability, and transparent risk decisions.

4.3 Multimodal Fusion and Anomaly Detection to Increase Resilience

Biometric systems based on a single modality often create single points of failure: they perform reliably under normal conditions but become vulnerable when a sensor is spoofed, degraded, or unavailable. Multimodal fusion mitigates this fragility by combining evidence from multiple modalities -such as face, voice, fingerprint, keystroke dynamics, gait, and contextual signals- to produce a more robust and discriminative identity estimate (Ross & Jain, 2004). Fusion can occur at the feature level (early fusion), score level (mid-level fusion), or decision level (late fusion), each offering different trade-offs in terms of transparency, computational cost, and response time. In adversarial settings, multimodal systems significantly increase attacker workload, as spoofing several modalities simultaneously -often under active liveness or presentation-attack-detection (PAD) controls- becomes far more challenging.

In parallel, anomaly detection introduces a complementary layer of protection by focusing not on identity confirmation alone, but on detecting deviations from typical behavioural or interaction patterns. This approach is especially well suited to continuous authentication, where behaviour naturally unfolds as a sequence rather than a single event (Gholami, Alaca, & Zulkernine, 2025). Techniques such as one-class classifiers, autoencoders, isolation forests, and density-based models allow systems to learn a user's characteristic manifold and to flag activity that falls outside expected bounds. Such mechanisms can detect impostors who might bypass a single check, identify legitimate users whose behaviour has been altered due to coercion or malware, and guide risk-adaptive responses that restrict or escalate access when anomaly scores exceed thresholds.

When viewed through a Zero-Trust lens, the integration of multimodal fusion with anomaly detection creates a layered and adaptive defence posture. Evidence becomes redundant (tolerant of sensor failure), diverse (harder to spoof), and temporal (evaluated continuously rather than once). This enriched evidential base enables finer-grained policy decisions—for example, allowing reduced-privilege access under moderate uncertainty, triggering step-up authentication under elevated risk, or isolating a session entirely when anomalous behaviour persists. Taken together, these capabilities shift the system from static recognition towards situational inference, where identity, context, and behavioural consistency are assessed continuously to maintain an appropriate and dynamically calibrated level of trust.

4.4 Generative Artificial Intelligence: Opportunities and Risks

Generative AI (GenAI) introduces strong advantages for biometric security but also creates significant new attack vectors. On the opportunity side, GenAI enhances simulation and red-teaming, enabling synthetic attack artefacts -such as altered illumination, voice perturbations, or spoofing materials (Khan and Khan, 2025)- to be generated for controlled robustness testing. It also supports data augmentation and domain randomisation, enriching training sets with rare poses, sensor noise patterns, or environmental variations, and improves rare-event modelling by synthesising edge cases that would otherwise be difficult to capture. Together, these capabilities help systems train against synthetic attacks and improve resilience without relying solely on large, hard-to-collect real-world datasets.

At the same time, GenAI significantly strengthens adversarial capabilities. Attackers can produce high-fidelity deepfakes, synthetic behavioural traces, and automated content that scales social-engineering efforts and enables synthetic identity fraud. This shortens the cycle between new attack techniques and their real-world deployment, expanding the overall attack surface.

Effective use of GenAI therefore requires robust presentation-attack detection, media-provenance checks, and challenge–response mechanisms that exploit signals difficult to synthesise. Continuous model hardening -through adversarial training and post-deployment monitoring- is essential. Beyond technical measures, sound governance is needed, including dataset lineage tracking, synthetic-data labelling, and clear documentation such as model cards and evaluation reports to ensure that GenAI contributes safely to the security stack.

In conclusion, AI does not improve biometrics by making them stricter. Instead, AI improves biometrics by making them continuous, contextual and probabilistic. Security transitions from being a verdict into being an ongoing assessment.

5. Zero Trust Architecture and the Integration of Biometrics

At this point, the role of biometrics in Zero Trust architectures is redefined from static trust anchors to dynamic, continuously evaluated signals. It outlines the core principles of Zero Trust, identifies optimal integration points for biometric verification, and demonstrates how adaptive trust models counter AI-driven identity threats in distributed environments.

5.1 The Principles of Zero Trust Architectures

Zero Trust -as already mentioned-, focuses on the basic principles of continuous verification, least privilege, and the elimination of implicit trust. Identities, devices, networks, and workloads are treated as untrustworthy by default. Access is granted only when needed and to an appropriate extent, based on dynamic, context-sensitive risks. Verification is event- and time-driven, policies reevaluate signals (identity status, device status, network context, behavioral telemetry) before access and throughout the session. Implicit trust (e.g., “inside the perimeter,” “recently authenticated,” or “biometric match == secure”) is removed; Instead, policy engines -so called policy decision points (PDPs)- create repeatable, auditable decisions that policy enforcement points (PEPs) enforce at every boundary. The architecture emphasizes deep defense (segmentation, strong identity, hardened endpoints, application-level controls) and telemetry fusion to detect anomalies and adjust permissions in real time. Crucially, Zero Trust decouples identity from the duration of permission: authentication is not a permanent pass, but a revocable, continuously assessed status. This mindset enables resilient responses to AI-powered threats, session hijacking, and synthetic identities in distributed environments.

5.2 Integration of Biometric Data into ZTA application

A coherent integration of biometrics into Zero Trust Architectures requires that biometric signals be treated as dynamic, context-dependent inputs within a continuously evaluated trust calculation, rather than as static authentication anchors. Their role only becomes effective when they are embedded in the policy loop -where PDPs coordinate decisions and PEPs enforce them- along with device status, behavioural telemetry, and cryptographically verifiable login data:

1. Step-up at sensitive boundaries (PDP-triggered): Biometric verification should be applied selectively when the PDP detects elevated risk conditions such as privilege escalation, anomalous device configuration, suspicious geolocation, or behavioural anomalies. In these situations, the PDP issues a step-up requirement that mandates new biometric verification with robust liveness and anti-spoofing guarantees. Biometric evidence serves as one of several confirming signals, including device attestations, possession-based evidence, and contextual risk assessments.

2. Continuous, seamless session assurance (PEP-enforced): During active sessions, PEP can maintain identity assurance through passive behavioural please biometrics and seamless indicators of user continuity. By periodically reevaluating policies—either time-based or event-driven—PEP can detect deviations or anomalous activity. Active biometric checks are only triggered when passive signals indicate increasing uncertainty, reducing the duration of session takeovers without creating constant friction points.
3. Account recovery and re-binding of identity: To prevent takeover through compromised recovery channels, re-binding workflows should combine hardware-based possession factors (e.g., FIDO2/WebAuthn) with multimodal biometrics and PAD-verified, trusted capture. This ensures that recovery cannot be undermined by synthetic identities or injected biometric artifacts.
4. High-risk remote onboarding (eKYC): In the onboarding context, biometrics complements cryptographic document verification with liveness detection, trusted device capture, and anti-injection attestations. Templates should be protected by revocable or transformation-based schemas, with clearly defined mode switching policies to transition users to alternative modalities when compromise is suspected.

5.3 Technical and Legal Requirements for the Processing and Storage of Biometric Data

Given the sensitivity and inherent irreversibility of biometric identifiers, their processing and storage must adhere to strict technical and regulatory requirements. From a technical standpoint, biometric pipelines require trusted capture, strong PAD, and protection against injection and replay, ensuring that only authentic signals enter the decision flow. Templates must be secured through cancellable or transform-based protection schemes that reduce the impact of compromise and prevent template inversion. Storage environments should use hardware-rooted isolation, encryption in transit and at rest, and strict segmentation to limit correlation across systems.

Legally, biometric data constitutes high-risk, special-category information, requiring explicit consent, purpose limitation, data minimization, and well-defined retention policies. Privacy-by-design measures -including transparent processing, documented risk assessments, and enforceable user rights-, are essential to ensure that biometric systems remain compliant, auditable, and proportionate. Together, these requirements establish the baseline for safe deployment within adaptive, risk-oriented identity architectures

6. Conclusion and Future Work

This work shows that the most consequential change in biometric authentication is architectural rather than a matter of squeezing out more recognition accuracy. In concert, AI methods and Zero-Trust principles shift authentication from a static, one-off judgement to a continuous, contextual, and probabilistic process. Systems engineered for stable signals, trusted environments, and cooperative users cannot meet the demands of cloud-native, distributed, and adversarial settings shaped by AI-enabled threats. In such environments, single checks and fixed thresholds are no longer adequate; what is required is a pipeline accumulating evidence over time, recalibrates confidence as conditions change, and ties those updates to clear policy actions.

Therefore, AI algorithms provides the technical substrate for this shift. Machine Learning supports temporal modelling of identity and per-user calibration, strengthening robustness as behaviour and operating conditions evolve. Multimodal fusion mitigates single-channel fragility by combining complementary signals, while Anomaly Detection reframes security as situational inference, surfacing deviations that simple matches miss. At the same time, Generative AI lowers the cost of high-fidelity spoofing and synthetic identities, exposing the limits of instance-based trust and expanding the attack surface. The result is an arms race demanding for continuous hardening, calibrated uncertainty, and policy-driven responses rather than static acceptance criteria.

Zero-Trust Architectures supply the operational fabric for these capabilities, wherein biometrics cease to function as immutable anchors and instead act as dynamic, policy-bound signals within a broader risk calculus -combined with device posture, network context, workload sensitivity, and behavioural telemetry. Trust becomes scoped, revisable, and revocable; policy decision points can step up, sustain, or curtail access as evidence evolves across the session.

Accordingly, any progress hinges on new evaluation criteria. Beyond headline accuracy, systems must demonstrate operational robustness to presentation and injection attacks, maintain tolerance to drift, and degrade gracefully when conditions deteriorate. They must also support revocability through cancellable templates, modality pivots, and secure re-binding processes. In parallel, privacy-by-design principles need to be embedded throughout the lifecycle, and biometric components must integrate cleanly into Zero-Trust

enforcement rather than operating as isolated mechanisms. Governance plays an essential role in making these properties transparent and auditable. Clear model cards, calibration reports, provenance tracking, and incident playbooks should enable independent verification of system behaviour and support accountable deployment. Looking ahead, the agenda is clear. Research and practice must standardise robustness benchmarks under generative-AI-driven threats, develop calibrated and explainable risk-scoring mechanisms, scale multimodal fusion with temporal anomaly modelling, and create practical revocation semantics that recognise the permanence of biometric data. Framed this way, AI-driven biometrics can contribute meaningfully to trustworthy digital identity - not as a single point of truth, but as a composable signal within a resilient, risk-based security ecosystem.

Acknowledgement

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 16KIS2239K SUSTAINET_guarDian). The authors alone are responsible for the content of the paper.

Ethics and Data Use Statement: This research used publicly available, anonymized secondary data. No identifiable personal information was accessed. Ethical approval was not required in accordance with institutional policies.

AI Assistance Statement: The authors used generative AI tools solely for linguistic editing and improving readability (e.g., grammar, phrasing, and structure). No AI system contributed to the formulation of the research questions, methodology, data interpretation, theoretical reasoning, or conclusions. All content was critically reviewed and approved by the authors, who take full responsibility for the manuscript.

References

Amirir, Z., Heidari, A., Jafari, N., and Hosseinzadeh, M., „Deep study on autonomous learning techniques for complex pattern recognition in interconnected information systems“, *Computer Science Review*, vol. 54, 2024, ISSN: 1574-0137, DOI: 10.1016/j.cosrev.2024.100666

Bengio, Y., Courville, A., Vincent, P., „Representation Learning: A Review and New Perspectives”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798 – 1828, 2013, DOI: 10.1109/TPAMI.2013.50

Buriro, A., and Luccio, F., “Mobile Biometrics: Innovations, Challenges, and Emerging Trends”, *Advanced Information Networking and Applications. AINA 2025. Lecture Notes on Data Engineering and Communications Technologies*, vol 252. Springer, Cham., DOI: 10.1007/978-3-031-87784-1_16

Farooq, M.U., Khan, A., Uddin, K., and Malik, K.M., “Transferable Adversarial Attacks on Audio Deepfake Detection”, *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, Tucson, AZ, USA, 2025, DOI: 10.1109/WACVW65960.2025.00178

Gholami, A., Alaca, F., and Zulkernine, M., “An anomaly detection based approach for continuous authentication with smartwatch inertial sensors”, *Computer & Security,* vol. 159.,2025, DOI: 10.1016/j.cose.2025.104656

Herbst, J., Rüb, M., Sanon, S.P., Lipps, C., and Schottem, H.D., “Medical Data in Wireless Body Area Networks: Device Authentication Techniques and Threat Mitigation Strategies Based on a Token-Based Communication Approach”, *network*, vol. 4, no. 2, 2024, DOI: 10.3390/network4020007

Jain, A.K., Nandakumar, K., and Ross, A., “50 years of biometric research: Accomplishments, challenges and opportunities”, *Pattern Recognition Letters,* Elsevier, 2016, DOI: 10.1016/j.patrec.2015.12.013

Kahn, F. A., and Kahn, M.K., „ Generative AI and Deepfake Detection in Biometric Systems“, *Cognitive Computation*, vol. 17, no. 112, pp. 1—21, 2025, DOI: 10.1007/s12559-025-10469-3

Kim, S., Tan,, Y.K., Jeong, B., Mondal, S., Aung, K.M.M., and Seo, J.H., “Scores Tell Everything about Bob: Non-adaptive Face Reconstruction on Face Recognition Systems”, *IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA, 2024, DOI: 10.1109/SP54263.2024.00161

Lipps, C., Herbst, J., and Schotten, H.D. “How to Dance your Passwords: A Biometric MFA-scheme for Identification and Authentication of Individuals in IIoT Environments”, *16th International Conference on Cyber Warfare and Security (ICCWS),* Cookeville, Tennessee, US, 2021.

Lipps, C., and Schotten, H.D., “Physical Layer Security: About Humans, Machines and the Transmission Channel”, *European Conference on Cyber Warfare and Security (ECCWS),* Chester, UK, 2022, DOI: 10.34190/eccws.21.1.403

Lipps, C., Bergkemper, L., Herbst, J., and Schotten, H.D., „ I Know You by Heart: Biometric Authentication based on Electrocardiogram (ECG) signals “, International Conference on Cyber Warfare and Security (ICCWS), Albany, NY, USA, 2022, DOI: 10.34190/iccws.17.1.12

Lipps, C., Reddy, R., Rüb, M., and Schotten, H.D., “Authentication in a Hyperconnected World: Challenges, Opportunities and Approaches”, *International Conference on Cyber Warfare and Security (ICCWS),* Johannesburg, South Africa, 2024, DOI: 10.34190/iccws.19.1.2070

Rose, S., Borchert, O., Mitchell, S., and Connelly, S., “Zero Trust Architecture”, NIST Special Publication 800-207, pp. 2—3, 2020, DOI: 10.6028/NIST.SP.800-207

Ross, A., and Jain, A.K., "Multimodal biometrics: An overview", 12th European Signal Processing Conference, Vienna, Austria, pp. 1221—1224, 2004, Print ISBN:978-320-0001-65-7

Ryu, R., Yeom, S., Herbert, D., and Dermoudy, J., „ The design and evaluation of adaptive biometric authentication systems: Current status, challenges and future direction", *ICT Express*, vol. 9, no. 6, pp. 1183—1197, 2023, DOI: 10.1016/j.icte.2023.04.003

Shukur, F., "Using Multiagents for Context-Aware Adaptive Biometrics", *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 5209—5210, 2017, DOI: 10.24963/ijcai.2017/764

Uppal, S., Banga, V., Neeraj, S., and Singhal, A., "A Comprehensive Study on Mitigating Synthetic Identity Threats Using Deepfake Detection Mechanisms", *14th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India, 2024, DOI: 10.1109/Confluence60223.2024.10463324

Verman, K., Kumar, R., Rao, A.P., and Ranjan, R., "Efficient e-KYC Authentication System: Redefining Customer Verification in Digital Banking", 9th International Conference on Signal Processing and Communication (ICSC), NOIDA, India, 2023, DOI: 10.1109/ICSC60394.2023.10441596

Villegas, W. E., Gutierrez, R., Navarro, A. M., and Mera-Navarrete, A., "Adaptive Authentication and Access Control System in Dynamic Educational Environments Based on AI", *Computer*, vol. 58, pp. 53—63, 2025, DOI: 10.1109/MC.2025.3565149

Wayman, J., Jain, A., Maltoni, D., and Maio, D., "Biometric Systems – Technology, Design and Performance Evaluation", Springer London, 2025, ISBN: 978-1-85233-596-0, DOI: 10.1007/b138151

AI Agents vs. Human Investigators: Balancing Automation, Security, and Expertise in Cyber Forensic Analysis

Sneha Sudhakaran[1] and Naresh Kshetri[2]
[1]Florida Institute of Technology, Melbourne, USA
[2]Rochester Institute of technology, Rochester, USA

ssudhakaran@fit.edu
naresh.kshetri@rit.edu

Abstract: In an era where cyber threats are rapidly evolving, the reliability of cyber forensic analysis has become increasingly critical for effective digital investigations and cybersecurity responses. Artificial Intelligence (AI) agents are being adopted across digital forensic practices due to their ability to automate processes such as anomaly detection, evidence classification, and behavioral pattern recognition, significantly enhancing scalability and reducing investigation timelines. However, the characteristics that make AI indispensable also introduce notable risks. AI systems, often trained on biased or incomplete datasets, can produce misleading results, including false positives and false negatives, thereby jeopardizing the integrity of forensic investigations. Furthermore, AI agents typically lack the contextual comprehension and ethical judgment required to interpret nuanced or legally sensitive scenarios. This study presents a meticulous comparative analysis of the effectiveness of the most used AI agent, ChatGPT, and human forensic investigators in the realm of cyber forensic analysis. Our research reveals critical limitations within AI-driven approaches, demonstrating scenarios in which sophisticated or novel cyber threats remain undetected due to the rigid pattern-based nature of AI systems. Conversely, our analysis highlights the crucial role that human forensic investigators play in mitigating these risks. Through adaptive decision-making, ethical reasoning, and contextual understanding, human investigators effectively identify subtle anomalies and threats that may evade automated detection systems. To reinforce our findings, we conducted comprehensive reliability testing of forensic techniques using multiple cyber threat scenarios. These tests confirmed that while AI agents significantly improve the efficiency of routine analyses, human oversight remains crucial in ensuring accuracy and comprehensiveness of the results. Our work validates the need for a hybrid forensic framework that combines the strengths of both AI automation and human expertise. Our study concludes by advocating for an integrated forensic analysis approach, proposing targeted strategies to incorporate both AI-driven efficiencies and human analytical insights. This collaborative model enhances overall forensic reliability, ensuring robust outcomes in the face of increasingly sophisticated cyber threats.

Keywords: AI agents, Cyber forensic analysis, Digital investigation, Human forensic analyst, Reliability

1. Introduction

In a recent report, cybercrime was estimated to cause damages of nearly $6 trillion globally by 2021, underscoring the scale and urgency of digital security challenges (Cybersecurity Ventures, 2020). To address this rise in cyber threats, digital forensics has emerged as a critical discipline that enables investigators to reconstruct events, identify malicious actors, and preserve evidence for legal proceedings (Casey, 2011). Among the evolving approaches in digital forensics, Artificial Intelligence (AI) has gained significant traction due to its ability to automate anomaly detection, evidence classification, and behavioral pattern recognition, thereby improving scalability and reducing investigation timelines (Garfinkel, 2010; Lillis et al., 2016).

However, the reliance on AI introduces notable risks. AI-driven forensic agents are often trained on incomplete or biased datasets, which can result in false positives or negatives that may compromise the integrity of investigations (Brennan & Luszcz, 2020). Moreover, AI lacks the contextual comprehension and ethical reasoning necessary to handle nuanced or legally sensitive scenarios—capabilities that remain the domain of human investigators (Boden, 2018). Human experts play an indispensable role in adapting to novel cyber threats, interpreting complex patterns, and applying professional judgment where AI systems fall short (James et al., 2014).

This study presents a comparative analysis of AI agents and human forensic investigators, highlighting the limitations of current AI-based approaches while demonstrating the enduring importance of human expertise. To reinforce our findings, we conducted comprehensive reliability testing across diverse cyber threat scenarios, evaluating both AI and human-centered forensic methods. Our results validate the need for a hybrid framework that integrates AI automation with human oversight to ensure accuracy, comprehensiveness, and reliability in digital investigations.

Contribution

- Examining the comparative strengths and weaknesses of AI agents and human investigators.

- Quantifying the risks of false positives/negatives in AI-based analysis and their impact on forensic integrity.
- Proposing a hybrid forensic framework that leverages AI efficiencies while retaining critical human oversight for contextual accuracy.

The rest of the paper is organized as follows: Section 2 presents the Background of this paper; Section 3 provides an overview of our Design and implementation; Section 4 presents the Evaluation of the proposed approach; Section 5 summarizes the Related Literature; and finally, section 6 presents the Conclusion.

2. Background and Related Work

Artificial Intelligence in Cyber Forensic Analysis - In recent years, Artificial Intelligence (AI) and Machine Learning (ML) have become central to advancing cyber forensic analysis by addressing the challenges posed by large-scale data volumes, speed, and the complexity of modern threats. For example, some researchers conducted a comprehensive analysis of how AI/ML techniques are being integrated into digital forensics and incident response, covering domains such as data collection and recovery, timeline reconstruction, pattern recognition, and managing the chain of custody. Similarly, in "Enhancing Cyber Forensics with AI and Machine Learning" Fakiha et al. (2023) show through case studies and surveys that organizations using AI have seen marked improvements in efficiency and precision, especially in threat detection and classification workflows. Another literature survey, "Artificial Intelligence in Digital Forensics: A Review of Cyber," highlights the growing need for trustworthy, scalable, and legally sound AI-based forensics systems, pointing out key challenges like explainability, adversarial robustness, and evidence admissibility. Explainable AI is also emerging as a crucial requirement; for instance, Alam & Altiparmak (2024) in "XAI-CF" argue that AI systems used in cyber forensics must be interpretable and transparent to be trusted by analysts and courts. Overall, while literature shows strong promise for AI to enhance speed, capacity, and pattern detection in forensic investigations, it also emphasizes the need for ethical frameworks, legal rigor, and human oversight.

AI-driven versus AI-assisted cyber forensic investigation- In recent literature, researchers distinguish clearly between AI-driven and AI-assisted forensic investigation approaches, highlighting trade-offs in autonomy, accuracy, and human oversight. An AI-driven approach refers to systems where much of the decision-making, pattern detection, or anomaly identification is automated, with minimal human intervention — for example, using machine learning or deep learning to flag suspicious activity, perform classification of threat artifacts, or even generate hypotheses of attack vectors. One study on financial cybersecurity investigations found that AI-driven techniques (e.g., anomaly detection, predictive modeling) substantially reduce investigation time and improve initial detection, but raise serious concerns regarding transparency, bias, and legal admissibility of evidence. In contrast, AI-assisted investigations retain humans more centrally, using AI tools as augmentation: assisting in triage, prioritizing artifacts, visualizing evidence, or suggesting leads rather than making final determinations. Fakiha et al. (2023) show that while AI/ML-assisted tools help forensic investigators by automating classification and threat detection tasks, ultimate judgments about evidence validity and context still depend heavily on expert human analysts. Similarly, the review "Artificial Intelligence in Digital Forensics: A Review of Cyber-Attack Detection Models and Frameworks" (2025) categorizes systems into fully automated (AI-driven), hybrid, and assisted models — noting that hybrid or assisted models tend to offer better legal defensibility and reduce risks of false positives/negatives because human analysts verify AI outputs. Moreover, the work XAI-CF (Explainable AI for Cyber Forensics) (Alam, 2024) emphasizes the importance of interpretability, arguing that AI-assisted (or hybrid) systems that incorporate explainability allow for greater trust from stakeholders such as courts and forensic examiners than fully automated black-box style AI-driven systems.

Complications on human users for AI versus human cyber forensic analysis - Human users encounter various challenges when comparing fully AI-driven forensic analysis with human-led or AI-assisted approaches. One major issue is trust and interpretability: many AI models, especially deep learning systems, are "black boxes" whose decision-making processes are opaque, making it difficult for investigators, legal professionals, or juries to understand how the AI arrived at a conclusion. This opacity can undermine confidence in the evidence or lead to challenges in court over admissibility (Alam and Altiparmak,2024;Hall, 2022; Anand, Thakur, 2025). Another complication involves bias and data quality: AI models trained on unrepresentative or biased datasets may amplify existing social or demographic biases — for example in predictive policing or automated recognition — resulting in unfair or erroneous outcomes that human overseers must attempt to correct or mitigate (Jinad et al., 2024; Solanke et al., 2022). There is also the problem of false positives/false negatives and robustness to adversarial conditions: AI tools may misclassify or miss evidence under conditions they were not trained for, whereas human experts may be better at noticing context, nuance, or unexpected irregularities (Casu et al.,

2023). Additionally, ethical and legal concerns arise, such as accountability — who is responsible if an AI-driven decision leads to a wrongful outcome — and privacy, especially when AI analyzes large amounts of sensitive data (Challenges and Limitations of AI in Forensic Science, 2025; Forensic Science Journal of Research, 2025). Finally, there are usability and human cognitive load issues: forensic practitioners must review, verify, and sometimes override AI outputs, while keeping proficiency with traditional methods and adapting to evolving AI tools — creating additional training and workload challenges that complicate adoption in practice (Solanke et al., 2022; Jinad et al., 2024)

Motivation of research- The rapid evolution of cyber threats has created an urgent demand for forensic methodologies that are both efficient and reliable. While Artificial Intelligence (AI) agents have demonstrated substantial potential in automating repetitive forensic tasks such as anomaly detection, evidence classification, and behavioral pattern recognition, their reliance on data-driven pattern matching exposes significant weaknesses. As the literature highlights, AI systems are often opaque, prone to bias, and vulnerable to adversarial conditions, resulting in false positives and false negatives that compromise the reliability and admissibility of forensic evidence (Hall, 2022; Jinad et al., 2024). Conversely, human investigators provide contextual judgment, adaptability, and ethical reasoning, but they face limitations in scalability and speed when confronted with the sheer volume and complexity of digital evidence (Solanke et al., 2022). These contrasting strengths and weaknesses underscore the need for a deeper comparative evaluation of AI-driven and human-centered forensic practices. A systematic analysis of the risks associated with false classifications in AI systems, and their impact on the integrity of forensic outcomes, is essential to guide their responsible adoption. At the same time, recognizing the irreplaceable role of human expertise motivates the development of frameworks that do not seek to replace human investigators but instead amplify their capabilities. This research is motivated by the vision of a hybrid forensic framework that combines the computational efficiency of AI with the contextual accuracy of human oversight. By quantifying the risks inherent in AI-based forensic analysis and systematically examining scenarios where human involvement remains indispensable, this work aims to establish a foundation for integrating AI and human intelligence based on evidence. Such a framework has the potential to enhance forensic reliability, strengthen evidentiary integrity, and ensure that digital investigations remain robust in the face of increasingly sophisticated cyber threats.

Table 1: Summary of Background, Related Works, and their Insights

Source	Related Work	Insight 1	Insight 2	Insight 3
Solanke et al. (2022)	Digital forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques	Forensics analysis makes use of ML models' recognition and pattern detection capabilities.	Applications of instruments in digital forensics with strengths & weaknesses in judicial.	Three critical instruments necessary (presented) for development of sound machine-driven digital forensics.
Fakiha (2023)	Enhancing Cyber Forensics with AI and Machine Learning: A Study on Automated Threat Analysis and Classification	Investigates the utilization of ML and AI in automated analysis & cyber threats classification.	Organizations that have implemented AI and ML in cyber forensics / threats with case studies.	Potential advantages of integrating AI and ML in advancing digital forensic investigations and their roles in cyber forensics.
Casu et al. (2023)	GenAI Mirage: the impostor bias and the deepfake detection challenge in the era of artificial illusions	Examines the impact of cognitive biases on decision making in digital forensics examination.	Novel "Imposter Bias" assuming generated by AI tools, multimedia contents like video, audio, and images.	Potential reasons and causes of Imposter Bias as it stems from an a priori assumption with impact to grow with increasing AI products.
Jinad et al. (2024)	Bias and fairness in software and automation tools in digital forensics	Analyze and discuss these biases present in software tools and automation systems used by law enforcement organizations and in court proceedings	Developed real-life cases and scenarios where some of these biases were identified that determined or influenced these cases	Research explains the increase validation in digital forensics software tools and ensure users' trust in the tools and automation techniques

3. Study Design

3.1 Research Design

This study adopts a **comparative experimental design** to evaluate the reliability of AI-driven versus human-centered cyber forensic investigations. The methodology emphasizes the identification of cases where AI produces misleading results, quantification of error rates, and assessment of the corrective role of human expertise.

3.2 DatasetSelection

We collected curated datasets from multiple domains, such as network intrusion traces, mobile and desktop memory dumps, malware binaries, and phishing email corpora. Each dataset will have pre-labeled "true" outcomes (verified by human experts and validated against benchmarks) to allow measurement of AI false positives and negatives.

3.3 AI Agent Deployment

We utilized multiple AI/ML forensic tools, covering anomaly detection (unsupervised models), classification (SVM, Random Forest, and Neural Networks), and pattern recognition (clustering and NLP models). AI systems will be provided with real-world forensic prompts and tasked with producing evidentiary findings.

3.4 Human Investigator Benchmark

A panel of experienced forensic researchers in our lab at Florida Institute of Technology independently examined the selected datasets. It ensured that we utilized all established forensic procedures, including timeline reconstruction, hash comparison, manual anomaly inspection, and evidentiary reporting.

3.5 Comparative Metrics

We compared the results from AI model analysis and human forensic researchers in a way that allows for the calculation and comparison of false positives and false negatives in AI outputs with human findings. Currently, we manually assessed our analysis results for validation as we tested this only on cases or prompts for analysis.

3.6 Hybrid Framework Testing

AI and human results will be combined using a hybrid workflow (AI for triage and clustering, humans for contextual validation). The performance metrics (accuracy, time efficiency, and evidentiary soundness) will be compared across AI-driven, human-led, and hybrid modes.

4. Case Study of Example Scenarios Demonstrating Untrustworthy AI Results

To empirically expose the weaknesses of AI-driven forensic tools, the study evaluates a set of carefully crafted forensic scenarios (Anand , Thakur, 2025). Each scenario highlights how an AI forensic agent produces results that deviate from human investigator analysis, emphasizing the risks of false positives, false negatives, and misinterpretation of context.

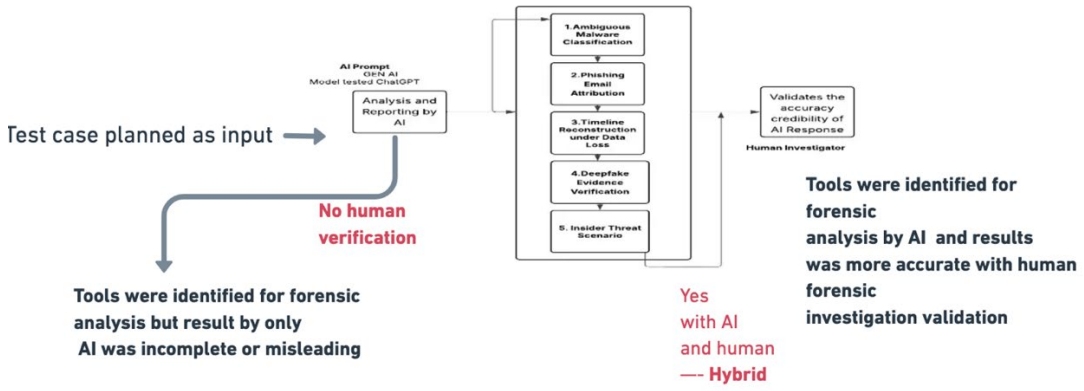


Figure 1: AI case study testing model for Ambiguity

The AI prompt is tested for various features in prompt analysis as shown in Figure1:

1. Ambiguous Malware Classification
2. Phishing email attribution
3. Timeline reconstruction
4. Deepfake Evidence
5. Insider threat

We simulated some AI prompts for testing the 5 above topics with ChatGPT and we recorded our interpretation of AI analysis below. The AI response of 15 pages was read and checked for authenticity by our forensic researchers in our lab to validate and our human forensic investigator interpretation is also written below.

4.1 Ambiguous Malware Classification

AI Prompt: Analyze this memory dump – 20 memory dumps with 10 benign and 10 infected were asked to analyze with basic AI prompt *"can you analyze the dump with volatility as a forensic investigator"* to identify all malware artifacts.
AI Agent Analysis: The AI model gives response with15-page report that flags multiple benign DLL injections and system processes as malicious, leading to inflated false positive rates. The response was very superficial.
Human Investigator Analysis: A forensic analyst cross-validates hashes against malware repositories and examines process lineage, correctly distinguishing between benign system injections and true malware artifacts.

4.2 Phishing Email Attribution

AI Prompt: *"Determine the attacker's identity based on the email header metadata."*
AI Agent Analysis: The AI tool extrapolates identity claims from spoofed headers, sometimes generating hallucinated or overconfident attributions that are not legally admissible. The output will point a lot of ambiguities in prediction of attacker and victim.
Human Investigator Analysis: An analyst recognizes spoofing techniques, verifies DomainKeys Identified Mail , Sender Policy Framework , Domain-based Message Authentication, Reporting & Conformance (DKIM/SPF/DMARC) inconsistencies, and reports that attacker attribution cannot be confirmed solely from headers, thereby maintaining evidentiary integrity.

4.3 Timeline Reconstruction under Data Loss

AI Prompt: *"Reconstruct the user activity timeline from partial Android memory images."*
AI Agent Analysis: The AI fills missing segments with interpolated or fabricated activities, producing timelines that appear complete but contain inaccuracies. Hallucinated output was point to unclear accusation.
Human Investigator Analysis: A human investigator marks the data gaps explicitly, acknowledges evidentiary incompleteness, and refrains from inserting non-existent activities, thus preserving reliability for court admissibility.

4.4 Deepfake Evidence Verification

AI Prompt: *"Verify if this video evidence is authentic."*
AI Agent Analysis: The AI model misclassifies an advanced deepfake as authentic due to limitations in its training dataset, producing a false negative. According to our analysis, this conclusion is more to be evaluated as Deepfake analysis is done only minimal until now
Human Investigator Analysis: The analyst applies multi-tool verification (frame analysis, metadata inspection, lighting/biometric consistency checks), and flags the video as suspicious, thereby reducing the chance of overlooking manipulated evidence.

4.5 Insider Threat Scenario

AI Prompt: We provided logs acquired from each devices (IOT, Unmanned Aeriel Vehicle, Android device, Desktop device)*"Identify insider threats from anomalous UAV, IOT, Android, Process Memory access logs."*
AI Agent Analysis: The AI system flags several legitimate after-hours activities (e.g., IT maintenance tasks) as malicious, generating high false positive rates.
Human Investigator Analysis: The human analyst correlates anomalies with schedules, support tickets, and contextual organizational knowledge, filtering legitimate activity from malicious behavior by using currently market avaliable tools in UAV, IoT, Desktop, Mobile forensic tools(Ahmed et al, 2016; Alam et al, 2024; Ali-Gombe et al, 2019; More et al, 2025),(Ali-Gombe, 2017),(Sudhakaran et al, 2022; Ligh et al, 2014 ; Sudhakaran et al, 2020; Sudhakaran,2022)

5. Example Scenarios of Empirical Success with Hybrid AI–Human Forensic Analysis

Table 2: Quantifying how we derived AI, Human led or Hybrid works better

Category	Scenario	Input Type / Size	Tools Used	Expected Outcome	Observed Outcome	AI-Driven	Human-Led	Hybrid
Benign	Windows 10 benign memory dump	RAW 1.5 GB	Vol3	Correct process listing	Achieved	Outlier not detect and hallucinated results		Yes
Malware	Windows 10 DLL injection	RAW 2 GB	Vol3 + malfind	Detect injected regions	Detected	Outlier not detect and hallucinated results		Yes
Malware	Win7 dump w/ keylogger	RAW 1 GB	Vol2 + Vol3	Identify hooks & keylogs	Partial detection	Outlier not detect and hallucinated results	Yes	Yes
Benign	Linux Ubuntu 22.04 memory	RAW 800 MB	Vol3 Linux plugins	Correct profile + pslist	Achieved	Outlier not detect and hallucinated results		Yes
Malware	Linux userland malware	RAW 1.2 GB	Vol3 + yara	Detect malicious ELF	Detected	Outlier not detect and hallucinated results		Yes
Benign	macOS Ventura (ARM64)	Snapshot 1.8 GB	Custom macOS tools	Extract processes	Achieved	Outlier not detect and hallucinated results		Yes
Benign	macOS Monterey (Intel)	Snapshot 1.6 GB	Volatility-mac	Browser artifacts	Achieved	Outlier not detect and hallucinated results		Yes
Benign	Android 11 memfetch	Dump 700 MB	memfetch + xxd	Extract app memory	Achieved	Outlier not detect and hallucinated results		Yes
Malware	Android 12 malware APK	Dump 900 MB	memfetch + xxd	Find malware artifacts	Detected			Yes
Benign	Android ARM64 emulator	Dump 650 MB	memfetch	Validate acquisition	Complete			Yes
Benign	Windows 11 benign	RAM 256 MB	Binwalk + strings	Recover credentials	Partially recovered	Outlier not detect and hallucinated results	Yes	Yes
Benign	Ubuntu 22.10 Linux	RAW 500 MB	Vol3 Linux	Parse Python services	Achieved			Yes
Benign	Windows 11 normal files	RAW 400 MB	dd, hexdump	Extract disk data	Achieved			Yes
Malware	Windows 11 corrupted files	RAW 300 MB	xxd, strings	Extract jpg/pdf files	Achieved		Yes	Yes
Malware	PCAP w/ C2 beaconing	PCAP 120 MB	Wireshark	Detect C2 periodicity	Detected	Outlier not detect and hallucinated results		Yes

Category	Scenario	Input Type / Size	Tools Used	Expected Outcome	Observed Outcome	AI-Driven	Human-Led	Hybrid
Malware	PCAP w/ DNS exfil	PCAP 90 MB	Zeek + Wireshark	Identify exfil queries	Detected			Yes
Malware	PCAP w/ TLS anomalies	PCAP 150 MB	Zeek	Flag abnormal JA3	Achieved			Yes
Malware	Memory + PCAP fusion	RAW + PCAP	Vol3 + Zeek	Correlate domains	Successful			Yes
Benign	Cross-OS Win + Linux analysis	RAW 3 GB	Vol3	Compare plugin behavior	Consistent			Yes
Malware	Android 8 malware	BIN 2 GB	Memfetch, LiME, DroidScraper, AmpleDroid	Evaluate reproducibility	High consistency	Outlier not detect and hallucinated results		Yes

We ran the same prompts with an AI-assisted analysis and human investigator, and the outcome was generated. The scenarios were extended to evaluate a **hybrid forensic framework** with light weight Dockerized containers (Fernalld et al, 2023)(O'Connor, 2024)used as template for analysis in which AI agents conduct the initial triage or pattern recognition, while human investigators validate, contextualize, and interpret the results. This combined workflow demonstrates how false positives and negatives can be reduced while preserving evidentiary integrity.

Ambiguous Malware Classification

- **Hybrid Outcome**: The AI agent rapidly identifies suspicious DLL injections, providing a candidate set of artifacts. The human investigator then validates these against known system libraries and malware repositories. False positives are filtered out, yielding both speed (via AI) and accuracy (via human validation).

Phishing Email Attribution

- **Hybrid Outcome**: AI tools extract header fields, cluster anomalies, and highlight possible spoofing indicators. Human investigators then verify these with DKIM/SPF/DMARC checks and organizational threat intelligence. Unlike AI-only analysis, attribution claims are tempered with caution to ensure legal defensibility.

Timeline Reconstruction under Data Loss

- **Hybrid Outcome**: The AI agent proposes a preliminary timeline by aligning recoverable memory fragments, while flagging gaps where interpolation would typically occur. Human investigators assess these gaps, annotate them as "missing evidence," and avoid overinterpretation. The result is a more transparent timeline with explicit markers of uncertainty.

Deepfake Evidence Verification

- **Hybrid Outcome**: AI models provide probabilistic scores on authenticity by scanning biometric inconsistencies and compression artifacts. Human investigators then conduct metadata verification and apply forensic video analysis. Together, this reduces the risk of false negatives, as humans can confirm suspicions that the AI might miss alone.

Insider Threat Scenario

- **Hybrid Outcome**: AI detects anomalous login times and access frequency, producing alerts that would otherwise overwhelm human analysts. Human investigators contextualize these alerts by cross-checking work schedules, maintenance logs, and personnel records. Legitimate anomalies are dismissed, leaving a refined set of truly suspicious activities.

Across all scenarios, the **hybrid model** demonstrates superior performance: AI contributes scale and efficiency in processing large datasets, while human investigators ensure contextual accuracy, interpretability, and legal

soundness. This combination directly addresses the limitations of both AI-driven and human-only methods, aligning with the study's contribution goals of comparative analysis, quantifying risks, and proposing a hybrid forensic framework. We would also like to educate students on how AI can be interlinked with future work for education and creating a correlation between cybersecurity and AI, similar to work done in (Suarez et al, 2024), Panakkadan et al, 2025) by consulting with actual forensic researchers with a vast set of AI prompts in multiple simulated case studies.

6. Conclusion

The increasing sophistication of cyber threats demands forensic approaches that are not only efficient but also reliable, interpretable, and legally defensible. This study has examined the comparative strengths and weaknesses of AI agents and human investigators, highlighting the risks posed by false positives, false negatives, and contextual blind spots when AI is deployed without oversight. While AI forensic agents excel at scalability and speed in tasks such as anomaly detection, evidence clustering, and timeline reconstruction, they often produce untrustworthy results when confronted with adversarial conditions, ambiguous data, or incomplete datasets. In contrast, human investigators provide adaptive judgment, contextual reasoning, and ethical decision-making, but are constrained by the scale and velocity of digital evidence. Through carefully designed scenarios—including ambiguous malware classification, phishing attribution, timeline reconstruction under data loss, deepfake verification, and insider threat detection—this research demonstrated how AI-only analysis can compromise evidentiary integrity, while human-only analysis, though reliable, struggles with efficiency. By quantifying error rates and examining case-specific discrepancies between AI and human findings, we showed that neither approach in isolation fully meets the demands of modern forensic investigations. The proposed hybrid forensic framework offers a pathway forward by integrating AI's computational efficiencies with the contextual accuracy of human oversight. In the hybrid model, AI provides rapid triage and pattern recognition, while human investigators validate results, resolve ambiguities, and preserve evidentiary soundness. This approach not only reduces error rates but also enhances interpretability, accountability, and admissibility of digital evidence. In conclusion, the contributions of this research are threefold: (1) a systematic comparative analysis of AI-driven and human-centered forensic investigations; (2) quantification of risks arising from false classifications in AI systems and their impact on forensic integrity; and (3) the design and validation of a hybrid framework that balances efficiency with contextual accuracy. Together, these contributions provide an evidence-based foundation for future forensic practice, ensuring that digital investigations remain robust, trustworthy, and resilient in the face of evolving cyber threats.

AI declaration: Artificial intelligence tools (OpenAI ChatGPT, GPT-5) were employed solely to improve grammar, readability, and formatting of the manuscript text. The authors take full responsibility for the scientific content, data interpretation, and conclusions. The authors also used this tool response as this work mainly focus on the usage of Gen AI tool like ChatGPT with Human forensic investigators for better efficacy over cyber forensic investigation. All outputs from AI tools were carefully reviewed and edited by the authors to ensure accuracy and originality.

Ethics Declaration: This study was conducted in accordance with established ethical and professional standards. No human participants, personal data, or sensitive information were involved; all datasets and experiments were performed on controlled test environments or publicly available resources. The work is intended solely for research and educational purposes in cybersecurity and does not promote or support malicious use.

References

Ahmed, I., Obermeier, S., Sudhakaran, S. & Roussev, V. (2017) 'Programmable logic controller forensics', *IEEE Security & Privacy*, 15(6), pp. 18–24.

Ahmed, I., Roussev, V., Johnson, W., Senthivel, S. & Sudhakaran, S. (2016) 'A SCADA system testbed for cybersecurity and forensic research and pedagogy', in *Proceedings of the 2nd Annual Industrial Control System Security Workshop*, pp. 1–9.

Alam, S. & Altiparmak, Z. (2024) 'XAI-CF – examining the role of explainable artificial intelligence in cyber forensics', *arXiv preprint*. Available at: https://arxiv.org/abs/... (Accessed: 30 September 2025).

Ali-Gombe, A., Sudhakaran, S., Case, A. & Richard III, G.G. (2019) 'DroidScraper: A tool for Android in-memory object recovery and reconstruction', in *Proceedings of the 22nd International Symposium on Research in Attacks, Intrusions and Defenses (RAID 2019),* pp. 547–559.

Ali-Gombe, A., Sudhakaran, S., Vijayakanthan, R. & Richard III, G.G. (2023) 'cRGB_Mem: At the intersection of memory forensics and machine learning', *Forensic Science International: Digital Investigation*.

Anand K.S, Thakur, S (2025) `Challenges and Limitations of AI in Forensic Science' (*International Journal of Research Publication and Reviews (IJRPR)*, 6(1), pp. 1–8. Available at: https://ijrpr.com/uploads/V6ISSUE1/IJRPR38264.pdf (Accessed: 30 September 2025).

Artificial Intelligence in Digital Forensics: A Review of Cyber (2025) *Journal of Information Systems Engineering and Management*.

Bhattathiripad, V.P., Sudhakaran, S. & Thalayaniyil, R.K. (2015) 'A new cyber forensic philosophy for digital watermarks in the context of copyright laws'.

Boden, M.A. (2018) *Artificial intelligence: A very short introduction*. Oxford: Oxford University Press.

Casey, E. (2011) *Digital evidence and computer crime: forensic science, computers, and the internet*. 3rd edn. Academic Press.

Casu, M., Guarnera, L., Caponnetto, P. & Battiato, S. (2023) 'GenAI Mirage: the impostor bias and the deepfake detection challenge in the era of artificial illusions', *arXiv preprint arXiv:2312.16220*. Available at: https://arxiv.org/abs/2312.16220 (Accessed: 30 September 2025).

Cybersecurity Ventures (2020) 'Cybercrime to cost the world $6 trillion annually by 2021'. Available at: https://cybersecurityventures.com (Accessed: 30 September 2025).

Fakiha, B. et al. (2023) 'Enhancing cyber forensics with AI and machine learning: a study on automated threat analysis and classification', *IET International Journal of Security and Sustainability Issues*.

Fernalld, K., O'Connor, T.J., Sudhakaran, S. & Nur, N. (2023) 'Lightweight symphony: towards reducing computer science student anxiety with standardized docker environments', in *Proceedings of the 24th Annual Conference on Information Technology Education*, pp. 15–21.

Forensic Science Journal of Research (2025) 'Developing an explainable AI system for digital forensics'. Available at: https://www.forensicscijournal.com/articles/jfsr-aid1089.php (Accessed: 30 September 2025).

Garfinkel, S.L. (2010) 'Digital forensics research: the next 10 years', *Digital Investigation*, 7(Suppl), pp. S64–S73.

Hall, S.W. (2022) 'XAI (explainable artificial intelligence) for digital forensics', *WIREs Forensic Science*, 4(5), e1434. doi: 10.1002/wfs2.1434.

James, J.I., Gladyshev, P. & Zhu, Y. (2014) 'Challenges with automation in digital forensic investigations', *Digital Investigation*, 11(3), pp. 204–214.

Jinad, R., Gupta, K., Simsek, E. & Zhou, B. (2024) 'Bias and fairness in software and automation tools in digital forensics', *Journal of Surveillance, Security and Safety*, 5, pp. 19–35.

Ligh, M.H., Case, A., Levy, J. and Walters, A., 2014. *The art of memory forensics: detecting malware and threats in windows, linux, and Mac memory*. John Wiley & Sons.

Lillis, D., Becker, B., O'Sullivan, T. & Scanlon, M. (2016) 'Current challenges and future research areas for digital forensic investigation', *ACM Computing Surveys*, 49(4), pp. 1–39.

More, S., Sudhakaran, S., O'Connor, T. & Carvalho, M. (2025) 'Comprehensive security assessment of Holy Stone drones: examining attack vectors', in *Proceedings of the International Conference on Cyber Warfare and Security*, pp. 574–583. Academic Conferences International Limited.

O'Connor, T.J., Schmith, A., Stricklan, C., Carvalho, M. & Sudhakaran, S. (2024) 'PWN lessons made easy with docker: toward an undergraduate vulnerability research cybersecurity class', in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education*, Vol. 1, pp. 986–992.

Panakkadan, R.R., Meher, P., More, S.A. & Sudhakaran, S. (2025) 'Enhancing cybersecurity education: the impact of simulated learning and interactive tutorials on student performance and anxiety reduction', in *INTED2025 Proceedings*, pp. 1775–1784. IATED.

Solanke, A.A. et al. (2022) 'Digital forensics AI: evaluating, standardizing and ensuring forensic soundness', *Künstliche Intelligenz*, 36, pp. 99–108. doi: 10.1007/s13218-022-00763-9.

Solanke, A.A. et al. (2022) 'Digital Forensics AI: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques', *Künstliche Intelligenz*, 36, pp. 143–161. doi: 10.1007/s13218-022-00763-9.

Suarez, L., Alshubrumi, D., O'Connor, T. & Sudhakaran, S. (2024) 'Unsafe at any bandwidth: towards understanding risk factors for ransomware in higher education', *Procedia Computer Science*, 238, pp. 815–820.

Sudhakaran, S. (2022) *Performing memory forensics for object recovery from Android application memory*.

Sudhakaran, S., Ali-Gombe, A., Case, A. & Richard III, G.G. (2022) 'Evaluating the reliability of Android userland memory forensics', in *Proceedings of the International Conference on Cyber Warfare and Security*, 17(1), pp. 423–432. Academic Conferences International Limited.

Sudhakaran, S., Ali-Gombe, A., Orgah, A., Case, A. & Richard, G.G. (2020) 'AmpleDroid: recovering large object files from Android application memory', in *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE.

Sudhakaran, S., Nur, N. & Zarraa, A. (2024) 'Navigating the challenges: understanding anxiety among international graduate students in the United States', in *INTED2024 Proceedings*, pp. 7664–7672. IATED.

Advancements in Developer-Focused IDE-Integrated Mobile App Security Scanning and Testing Tools

Franck Monga Tamala, Noluntu Mpekoa and Khutso Lebea
University of Johannesburg, South Africa

francktamala@gmail.com
noluntum@uj.ac.za
klebea@uj.ac.za

Abstract: The increasing number of mobile applications in modern society has significantly increased the possibility of security vulnerabilities, particularly as users increasingly rely on these applications for sensitive tasks such as banking, communication, and e-commerce. Any flaw within a mobile application may result in significant privacy violations, financial loss, or data exposure. The development of secure mobile applications presents a persistent challenge, especially for novice developers who often lack the expertise or tools to detect vulnerabilities during the early stages of coding. As mobile platforms become increasingly complex and threats become more sophisticated, integrating effective security practices into the software development lifecycle (SDLC) has become imperative. While a variety of security tools exist to support vulnerability detection, many fail to offer real-time, developer-friendly support embedded directly within Integrated Development Environments (IDEs), leaving a critical security gap, especially for novice developers. This paper conducts a systematic literature review on developer-focused tools available for real-time mobile app security scanning and IDE integration. The review emphasises tools that assist developers directly within Integrated Development Environments (IDEs), focusing on practical support during coding rather than post-deployment analysis. The objective was to thoroughly identify both the strengths and weaknesses of the existing tools that provide real-time mobile app security scanning. The PRISMA 2020 statement, which provides a comprehensive framework for conducting systematic literature reviews, served as the foundational guideline for this study. A thorough search was conducted to retrieve relevant journal articles and conference papers published between 2020 and 2025. This selection criterion ensured that the study incorporated the most recent and relevant findings in the field. Each identified publication was meticulously evaluated for its relevance, quality, and contribution to the existing body of knowledge, thereby enriching the systematic review process. The findings suggest that while existing tools contribute significantly to automation, benchmarking, privacy scanning, malware detection, and dependency management, they remain fragmented and largely external to developer workflows. Most require execution outside the IDE, lack lightweight integration, and fail to deliver real-time vulnerability feedback during coding. Even industry tools such as MobSF, NowSecure, and Checkmarx provide powerful analysis but operate as standalone platforms rather than IDE-embedded solutions. This gap is particularly critical in agile and novice development contexts, where immediate, contextualised security feedback is necessary to prevent vulnerabilities at their source.

Keywords: Mobile app security scanning, Mobile application vulnerability scanning, Real-Time scanning, IDE integration

1. Introduction

The increasing number of mobile applications in modern society has considerably amplified the potential for security vulnerabilities, especially as users increasingly depend on these applications for critical activities such as banking, communication, and e-commerce (Statista, 2025; National Institute of Standards and Technology (NIST), 2023). Any vulnerability in a mobile application can lead to serious privacy breaches, financial repercussions, or data exposure. The creation of secure mobile applications remains a significant challenge, particularly for less experienced developers who frequently do not possess the necessary knowledge, experience, or immediate tools to detect and address vulnerabilities in the initial phases of development (Peruma et al, 2024). With the increasing complexity of mobile platforms and the sophistication of security threats, it is essential to integrate proactive and developer-friendly security mechanisms into the Software Development Lifecycle (SDLC). This integration has emerged as a necessity and an industry best practice (OWASP, 2023).

While there is an array of security tools available, including static and dynamic analysis as well as hybrid systems, numerous options tend to be inadequate in one or more essential aspects. Frequent challenges encompass significant computational demands, complex technical interfaces, inadequate integration with widely-used development environments, and a lack of real-time feedback mechanisms (Chen et al, 2022; Salihu et al, 2019; NowSecure Inc., n.d.). Specifically, many tools function independently of the developer's main workflow, frequently necessitating post-build execution or the use of external dashboards (Feng et al, 2020; Gadient et al, 2019). This disconnect not only increases mental workload but also undermines the development of secure coding practices, especially for novice developers who may struggle to interpret delayed or ambiguous alerts (Haug, da Silva and Wagner, 2022). The absence of solutions integrated within Integrated Development

Environments (IDEs) presents a significant gap in modern mobile app development methodologies, ultimately undermining security results and prolonging the remediation process.

This study aims to address the existing gap by performing a systematic literature review on real-time mobile application security tools, with a particular emphasis on those that integrate with Integrated Development Environments (IDEs). The review synthesises recent academic research from 2020 to 2025, focusing on six key studies: Palma et al (2020), Mohsen, Oosterhaven and Turkmen (2021), Zhao et al (2022), Lemos et al (2022), Rahkema and Pfahl (2022), and Zhu et al (2024). In addition, widely used industry platforms such as MobSF, NowSecure, and Checkmarx are considered (Abraham, 2023; NowSecure Inc., n.d.; Checkmarx, 2024). The results indicate that although these tools demonstrate valuable contributions, such as automation of security testing, benchmarking of static analysis, privacy scanning, malware detection, and dependency management, most remain fragmented and external to developer workflows. None provide lightweight, real-time, and contextualised feedback directly within IDEs for Android development, with the exception of one tool designed for iOS dependency management (Rahkema and Pfahl, 2022), which is a critical requirement for agile development workflows and secure programming practices (Fu et al, 2024; Steenhock et al, 2024). This research holds considerable importance as it provides a timely assessment of the technological and usability challenges faced by developers, particularly in dynamic and resource-limited settings. This paper highlights the shortcomings of existing tools and the unaddressed needs of novice developers, thus contributing to the discourse on developer-centric security practices and suggesting practical options for future research and tool improvement (Peruma et al, 2024; Haug et al, 2022).

The remainder of this paper is structured as follows: Section 2: Research Methodology outlines the PRISMA-based approach used to select and evaluate relevant academic sources. Section 3: Study background presents the gap in the literature by looking at current studies. Section 4: Results and Findings present the key insights drawn from the selected studies. Section 5: Discussion interprets these findings in the context of developer workflows, tool adoption barriers, and security effectiveness. Section 6: Recommendations and Future Work proposes strategies for designing more usable, lightweight, and IDE-integrated security solutions. Section 7: Conclusion summarises the key contributions and calls for more developer-focused innovation in mobile app security tooling.

2. Research Methodology

This research utilised a Systematic Literature Review (SLR) to examine real-time mobile application security scanning tools, specifically emphasising those that facilitate static and dynamic analysis, integrate with IDEs, and provide feedback mechanisms designed for developers. In contrast to conventional literature reviews, a systematic literature review adheres to a strict, transparent, and replicable protocol aimed at minimising bias and guaranteeing thorough coverage of pertinent studies (Page et al, 2021). The PRISMA 2020 statement provided the essential framework for the planning, execution, and reporting of this examination.

2.1 Research Approach

The review was carried out leveraging two prominent digital libraries that are extensively referenced in the fields of software engineering and cybersecurity research: IEEE Xplore and the ACM Digital Library. The selection of these databases is based on their indexing of high-quality, peer-reviewed journals and conference proceedings, which represent the primary publication formats within the fields of computer science and engineering. A systematic search query was implemented utilising various combinations of the specified terms and Boolean operators:

- **IEEE:** mobile app security scanning OR mobile application vulnerability scanning AND real-time scanning OR IDE integration
- **ACM:** "mobile app security" OR "mobile application vulnerability" AND "real-time scanning" OR "live analysis" OR "IDE integration" AND "developer tool" OR "security framework"

The search was limited to publications from 2020 to 2025 to maintain relevance and recency. This six-year period highlights the latest developments in mobile security tools and illustrates the progression of developer-focused strategies for identifying vulnerabilities.

2.2 Criteria for Inclusion and Exclusion

The selection of studies was based on the following criteria:

- **Inclusion criteria:**

- *Peer-reviewed journal or conference papers.*
- *Published between 2020–2025.*
- *Explicitly addressed mobile app security scanning, static or dynamic analysis, or dependency-based vulnerability detection.*
- *Evaluated or proposed a tool/framework relevant to mobile application security.*
- **Exclusion criteria:**
- *Studies outside the 2020–2025 range.*
- *Grey literature (e.g., theses, whitepapers, blogs).*
- *Papers focusing exclusively on non-mobile domains (e.g., IoT, EV charging, healthcare apps, or generic cloud/cryptography approaches).*
- *Regulatory reviews or conceptual discussions without tool evaluation.*

2.3 Screening and Selection Process

The initial search resulted in 842 papers from both IEEE Xplore and the ACM Digital Library. Following the elimination of duplicates and the implementation of title and abstract screening, a total of 366 papers were retained. The assessment was conducted in detail according to the established eligibility criteria. From this selection, 6 (six) primary studies were identified as directly pertinent to the research objective. The selection of these 6 papers is based on their direct evaluation of security scanning tools for mobile applications or their introduction of innovative frameworks that enhance automated, developer-focused security practices.

2.4 Extraction and Analysis of Data

For every study included, the following data points were extracted:

- Identification of the tool's name along with its intended function, such as static analysis, dynamic analysis, or hybrid detection.
- Identify the target platform: Android, iOS, or cross-platform.
- Integration methodology (IDE plugin, app build scanner (e.g. APK), dependency verifier).
- Assessment techniques (benchmarks, CVEs, experimental datasets).
- Strengths and limitations include aspects such as real-time capability, false positive rate, and developer usability.

This structured comparison facilitated a synthesis across studies regarding the effectiveness of existing tools in addressing mobile app vulnerabilities while also highlighting the persistent gaps, especially the absence of real-time, IDE-integrated feedback mechanisms for developers.

Although more than 800 papers were initially retrieved from databases such as IEEE Xplore, ACM Digital Library, and ScienceDirect, the application of strict inclusion criteria substantially narrowed this number. Only studies that explicitly addressed real-time security scanning, IDE integration, or lightweight implementation in mobile application development were considered eligible. As a result, six primary studies met most of these criteria and were included in the final analysis. This limited number of studies underscores the emerging nature of this research area and the scarcity of targeted academic work focusing specifically on developer-orientated, IDE-integrated security solutions.

The following section discusses the background literature that contextualises the results of this review and highlights the persistent gap in IDE-integrated mobile security tools.

3. Study Background

This section provides the conceptual background and identifies the existing research gap that motivated the systematic review outlined in the preceding section. The security of mobile applications has emerged as a pressing concern, driven by their extensive adoption and increasing complexity. In light of the progress made in vulnerability detection and analysis, it is noteworthy that the majority of existing solutions function primarily post-deployment or within external testing environments (Palma et al, 2020; Zhao et al, 2022; Zhu et al, 2024; Lemos et al, 2022). There is a notable absence of lightweight, real-time, and IDE-integrated scanning tools that could effectively support developers throughout the coding process. This section critically examines recent literature, highlighting existing weaknesses, demonstrating support regarding the identified gap, and positioning the contribution of this study within the broader academic literature.

3.1 Literature Review – Identified Weaknesses

Palma et al (2020) introduced an automated security testing methodology for Android applications, illustrating that automated pipelines can enhance the speed of vulnerability detection in mobile software. Nonetheless, the tool operates outside the development environment and fails to deliver real-time alerts within the IDE, which consequently restricts its efficacy for iterative and secure coding practices. Mohsen, Oosterhaven, and Turkmen (2021) presented KotlinDetector, a tool aimed at detecting Kotlin within Android applications and assessing the potential security risks associated with mixed-language projects. Their research emphasised the security trade-offs associated with adopting Kotlin and showcased the effectiveness of binary analysis. However, KotlinDetector operates in conjunction with external vulnerability scanners like AndroBugs, rather than being integrated into the development workflow. This indicates that developers continue to depend on manual processes outside the IDE.

Zhu et al (2024) performed an extensive benchmarking analysis of static application security testing (SAST) tools specifically for the Android platform. The researchers assessed 11 tools utilising both synthetic and real-world Common Vulnerabilities and Exposures (CVE) datasets and introduced a unified platform, VulsTotal, aimed at standardising comparisons. Their study highlighted notable coverage gaps and inconsistencies among various tools, while also indicating that SAST continues to operate as an offline process, separate from real-time software development practices. On the other hand, Zhao et al (2022) introduced a framework for scanning privacy information through static analysis, integrating text analysis of privacy policies with code-level examination to detect inconsistencies. Their approach exhibited significant accuracy coupled with minimal overhead, thereby enhancing its efficiency for compliance objectives. Nonetheless, the tool lacks integration within IDEs, which results in developers being unable to receive immediate warnings during coding, thereby diminishing its preventative effectiveness.

Lemos et al (2022) investigated a hybrid approach to malware detection that integrates metadata with inter-process communication (IPC) analysis. The machine learning model demonstrated an accuracy rate of 87% in differentiating between malicious and benign applications. This approach, while promising for malware detection, necessitates the execution of applications and the collection of runtime data, which renders it impractical for providing real-time support to developers during the coding process. Another study by Rahkema and Pfahl (2022) introduced SwiftDependencyChecker, a tool that seamlessly integrates with Xcode to identify outdated or vulnerable dependencies specified via CocoaPods, Carthage, and Swift Package Manager. This constitutes one of the limited efforts towards direct integration with IDEs. However, its focus is limited to iOS ecosystems and dependency management, lacking the capability for comprehensive vulnerability scanning across Android or various types of vulnerabilities.

3.2 Literature Review – Consensus on the Gap

Collectively, these studies indicate an agreement that vulnerabilities in mobile applications can be effectively identified through static analysis, hybrid analysis, or dependency checks. Nonetheless, all six studies exist outside of IDE workflows or are limited in focus to particular threat vectors. Even in cases where IDE integration is available, such as with SwiftDependencyChecker (Rahkema et al, 2022), its applicability is limited and does not extend to Android. This body of work highlights a significant research gap: developers currently do not have access to automated, lightweight, real-time scanning solutions integrated within their IDEs that can proactively prevent vulnerabilities during the development phase, rather than addressing them post-deployment.

3.3 Contribution

This research provides a systematic review of cutting-edge tools and highlights the lack of comprehensive IDE-integrated solutions for mobile app development. This work builds upon this synthesis by highlighting the strengths and weaknesses of existing approaches, while also emphasising the necessity for tools that offer instantaneous, developer-friendly feedback without requiring users to exit the IDE. By clearly delineating this gap, this study establishes a basis for enhancing real-time, automated vulnerability scanning as a fundamental component of mobile application development.

4. Findings

This research adopted the PRISMA 2020 framework to conduct a systematic review of current tools for mobile application security. Six primary studies were identified (2020–2025), each presenting distinct methodologies for vulnerability detection. The primary aim was to identify real-time security scanning, IDE integration, and lightweight solutions. The findings indicate that although certain tools facilitate automated testing (Palma et al,

2020; Zhu et al, 2024) or offer integration with development ecosystems (Rahkema et al, 2022), there is currently no comprehensive real-time solution that is embedded within IDEs. Tools such as MobSF, NowSecure, and Checkmarx (Abraham, 2023; NowSecure Inc., n.d.; Checkmarx, 2024) remain widely used in practice but face the same limitation: they operate as external platforms, requiring post-coding scans instead of immediate feedback.

The second objective involved identifying the strengths and weaknesses of the previously identified solutions. Table 1 provides a summary of the key features, strengths, and limitations of the six tools that were reviewed.

Table 1: Strengths and weaknesses of existing tools

Tool / Study	Main Focus	Strengths	Weaknesses
Palma et al (2020)	Automated Android security testing	Pipeline automation	External to IDE; no real-time feedback
Mohsen et al (2021)	Kotlin usage & vulnerabilities	Language-specific detection	Needs external scanners; not IDE-integrated
Zhu et al (2024)	SAST benchmarking	Coverage insights; VulsTotal platform	Inconsistent detection; offline
Zhao et al (2022)	Privacy scanning (static + text)	High accuracy; low overhead	External compliance tool only
Lemos et al (2022)	Malware detection (metadata/IPC)	Strong runtime accuracy	Requires execution; no IDE integration
Rahkema & Pfahl (2022)	Swift dependency checker	Direct IDE integration (Xcode)	Narrow scope; iOS only

As shown in Table 1, the reviewed studies reveal that while existing tools exhibit distinct technical strengths, their implementation scope remains fragmented. One automated Android security testing pipelines to streamline post-build vulnerability detection, demonstrating the feasibility of continuous scanning within CI/CD environments. Another focused on Kotlin-specific vulnerabilities, enhancing language-awareness but operating entirely outside the IDE. Benchmarking work on Static Application Security Testing (SAST) tools exposed inconsistencies in vulnerability coverage, illustrating the variability of current methods. A separate study developed a privacy scanning framework that combined static and textual analysis to detect data leaks but lacked IDE integration. Another proposed malware classification using metadata and inter-process communication features, achieving high detection accuracy but again functioning externally. The final study demonstrated IDE-level integration through dependency management in iOS SwiftDependencyChecker (Rahkema et al, 2022), showing that integration is possible but currently platform-limited.

Collectively, these findings highlight both technical maturity and integration immaturity; the tools are effective in their respective niches but do not provide a seamless, real-time experience for developers within Android Studio or comparable IDEs.

5. Discussion

This research was initiated with two specific objectives. The first was to identify real-time security scanning, IDE integration, and lightweight solutions. The findings indicate that none of the six academic tools examined provides comprehensive IDE integration for Android development. Only SwiftDependencyChecker (Rahkema et al, 2022) demonstrates IDE embedding; however, its application is restricted to iOS and focused solely on dependency management. Other tools, such as KotlinDetector (Mohsen et al, 2021) and the VulsTotal benchmarking platform (Zhu et al, 2024), make important contributions in language-specific analysis and comparative evaluation of static analysis tools, yet both require execution outside of the developer environment. This absence of real-time integration mirrors the limitations of widely used industry platforms, including MobSF, NowSecure, and Checkmarx (Abraham, 2023; NowSecure Inc., n.d.; Checkmarx, 2024). While these tools are effective at detecting vulnerabilities post-build, they disrupt developer workflows by requiring

separate setup and execution, preventing them from providing immediate, contextualised feedback during coding.

The second objective was to identify the strengths and weaknesses of the solutions under review. Each of the six academic studies contributes valuable insights. Palma et al (2020) demonstrated the feasibility of automated pipelines for Android applications, showing how testing processes can be streamlined. Mohsen, Oosterhaven et al (2021) highlighted the growing importance of language-specific vulnerabilities with the adoption of Kotlin, addressing an area often overlooked in broader security analyses. Zhu et al (2024) provided systematic benchmarking that revealed inconsistencies in static analysis coverage, offering a clearer picture of how SAST tools perform against real-world CVEs. Zhao et al (2022) developed a framework that combined privacy policy analysis with static code scanning, achieving high accuracy in detecting compliance gaps. Lemos et al (2022) added a runtime perspective by demonstrating effective malware classification using metadata and inter-process communication features. Finally, Rahkema et al (2022) proved that IDE-level integration is technically feasible, even though it is limited in scope.

Despite these contributions, notable weaknesses persist across the literature. The first is a reliance on external scanning platforms, which undermines real-time developer support. The second is fragmented coverage, with each tool addressing only a subset of vulnerabilities, whether dependencies, privacy, language constructs, or runtime behaviour, rather than delivering a holistic solution. Third, there is an absence of lightweight integration with IDEs. Where integration does exist, it is limited to narrow contexts and ecosystems (e.g., iOS dependencies). Finally, there are issues of limited applicability, as most tools require significant computational resources, manual configuration, or external datasets, which can discourage adoption among novice developers.

These findings collectively highlight a significant research gap in the realm of real-time, IDE-integrated, lightweight security scanning for mobile app development. The gap is particularly consequential in agile development environments, where continuous integration and rapid iteration demand immediate, actionable feedback. For novice developers, the absence of IDE-integrated security support increases the risk of introducing vulnerabilities during the earliest stages of coding, when mistakes are most easily corrected but often overlooked. Thus, while current academic and industry tools provide valuable insights and targeted solutions, they remain insufficient to meet the practical needs of secure, real-time mobile app development.

6. Recommendations and Future Work

This study is important because it highlights the lack of proactive security tools that support developers during the actual coding process. While existing approaches contribute to automation, benchmarking, and detection accuracy, they remain reactive, often functioning after deployment or in external pipelines. This limits their ability to prevent vulnerabilities at the point of introduction. To advance the field, future research should build on the lessons learnt from both academic studies and industry platforms.

Three key lessons emerge from the strengths of existing tools.

- **Automation reduces developer workload.** Palma et al (2020) demonstrated that automated pipelines can accelerate vulnerability detection, showing that developers benefit when routine security checks are handled automatically rather than manually.
- **Lightweight static and hybrid analysis can achieve high accuracy.** Tools such as Zhao et al (2022) and Lemos et al (2022) highlight that carefully designed analysis frameworks can produce accurate results without imposing excessive computational costs. This suggests that resource-efficient solutions are feasible for integration into IDEs.
- **IDE integration is feasible in principle.** Rahkema et al (2022) proved that embedding security scanning directly into an IDE is technically possible, though limited in scope. This demonstrates a foundation that future solutions can extend to broader vulnerability coverage.

Building on these lessons, several recommendations emerge.

Firstly, future work should focus on the design of lightweight IDE plugins or extensions capable of real-time vulnerability scanning. Such tools should combine multiple analysis approaches: static analysis for code-level issues, hybrid analysis for runtime-like conditions, and dependency scanning for third-party libraries. By merging these capabilities, developers could receive comprehensive security feedback without leaving their development environment. Second, there is a strong need for usability-focused research. Tools that disrupt developer workflows are unlikely to be adopted, regardless of their detection accuracy. Future studies should therefore evaluate developer adoption and usability through empirical testing, with particular attention to

novice programmers. Novices often lack the expertise to interpret external scan results and would benefit greatly from contextualised, in-line feedback that explains vulnerabilities and provides actionable remediation guidance.

Secondly, future research should adapt lessons from industry platforms such as MobSF, NowSecure, and Checkmarx (Abraham, 2023; NowSecure Inc., n.d.; Checkmarx, 2024). These platforms already provide powerful scanning capabilities but function as standalone environments, requiring developers to upload applications or configure external systems. The challenge and opportunity lie in shifting these capabilities into IDE-integrated, on-the-fly analysis. By doing so, the efficiency of industry-grade tools could be combined with the immediacy required in agile and continuous integration environments. Fourth, cross-platform considerations should be addressed. While this study focused primarily on Android, iOS developers face similar challenges. SwiftDependencyChecker (Rahkema et al, 2022) illustrates a narrow but valuable step toward IDE integration. Future work should extend integration beyond a single ecosystem to provide consistent, secure development practices across platforms.

Finally, collaboration between academia and industry is essential. Academic prototypes often demonstrate innovation but lack scalability, while industry tools are robust but limited in scope. Future research should prioritise bridging this divide, combining rigorous academic methods with the resources and adoption pathways available to industry. In summary, while the challenges surrounding secure mobile app development are well documented, this study provides a unique synthesis by focusing specifically on developer-centric, IDE-integrated security solutions. By consolidating evidence from both academic and industry sources, it establishes a comprehensive foundation for future empirical research and the design of tools that deliver real-time, contextual vulnerability detection directly within development environments.

7. Conclusion

The aim of this paper was to evaluate whether existing mobile application security tools provide real-time, IDE-integrated, lightweight scanning. Using the PRISMA 2020 methodology, six academic tools were systematically reviewed, alongside industry platforms such as MobSF, NowSecure, and Checkmarx. The findings indicate that current solutions make several valuable contributions. Some tools advance automation through testing pipelines, others address language-specific vulnerabilities, benchmark static analysis performance, improve privacy detection, or enhance malware classification. One solution even demonstrates that IDE-level integration is technically feasible, though limited in scope. Similarly, industry platforms are powerful but operate externally. Despite these strengths, all solutions share a fundamental limitation: they are fragmented, reactive, and not embedded into Android IDEs for real-time feedback. The high-level conclusion is clear: the research gap remains unaddressed. Developers, especially novices, lack immediate security feedback during coding, and existing tools do not meet the demands of agile, continuous integration workflows. This study contributes by clearly articulating this gap and synthesising lessons from both academic and industry efforts. The findings emphasise that while advances have been made in automation, accuracy, and benchmarking, these remain insufficient to support secure development practices from the ground up. Moving forward, the development of lightweight, IDE-integrated, cross-platform solutions should be prioritised. Such tools would allow vulnerabilities to be identified and remediated at the earliest stage of the software lifecycle, reducing costs, improving security, and fostering more secure coding practices among novice and experienced developers alike.

AI Declaration: Artificial intelligence (AI) tools were used to support the preparation of this paper. AI was employed to assist with language refinement and structural editing. The author retained full responsibility for reviewing, verifying, summarising, and interpreting all outputs, ensuring the accuracy and integrity of the final paper.

Ethics Declaration: Ethical clearance was not required for this research, as it was based solely on a systematic review of published academic and industry literature. No human participants or sensitive data were involved in the study.

References

Abraham, A. (2023) *Mobile Security Framework (MobSF)*. Available at: https://github.com/MobSF/Mobile-Security-Framework-MobSF.

Checkmarx (2024) *Checkmarx Application Security Platform*. Available at: https://www.checkmarx.com.

Chen, S., Zhang, Y., Fan, L., Li, J. and Liu, Y. (2022) 'AUSERA: Automated Security Vulnerability Detection for Android Apps', *Proceedings of the IEEE/ACM International Conference on Software Engineering*, pp. 1120–1131.

Feng, R., Chen, S., Xie, X., Meng, G., Lin, S.-W. and Liu, Y. (2020) 'Limitations of Existing Mobile Security Tools', *ACM Transactions on Software Engineering and Methodology*, 29(4), pp. 1–24.

Fu, Y., Xu, Q., Hu, J., Wang, X. and Song, Z. (2024) 'AIBUGHUNTER: AI-Powered Real-Time Vulnerability Detection for IDEs', *IEEE Transactions on Software Engineering*. doi: 10.1109/TSE.2024.123456.

Gadient, P., Ghafari, M., Frischknecht, P. and Nierstrasz, O. (2019) 'Tool Support for Security in Agile Development', *Proceedings of the International Conference on Agile Software Development*, pp. 152–167.

Haug, M., da Silva, A.C.F. and Wagner, S. (2022) 'CRITICALMATE: Real-Time Security Feedback in IDEs', *Empirical Software Engineering*, 27(5), pp. 1–30.

Lemos, R., Heinrich, T., Maziero, C.A. and Will, N.C. (2022) 'Is It Safe? Identifying Malicious Apps Through the Use of Metadata and Inter-Process Communication', *2022 IEEE International Systems Conference (SysCon)*. doi: 10.1109/SysCon53536.2022.9773881.

Mohsen, F., Oosterhaven, L. and Turkmen, F. (2021) 'KotlinDetector: Towards Understanding the Implications of Using Kotlin in Android Applications', *2021 IEEE/ACM 8th International Conference on Mobile Software Engineering and Systems (MobileSoft)*. doi: 10.1109/MobileSoft52590.2021.00018.

National Institute of Standards and Technology (NIST) (2023) *National Vulnerability Database*. Available at: https://nvd.nist.gov (Accessed: 1 September 2025).

NowSecure Inc. (n.d.) *NowSecure Platform Overview*. Available at: https://www.nowsecure.com (Accessed: 1 September 2025).

OWASP (2023) *Mobile Application Security Verification Standard (MASVS)*. Available at: https://owasp.org/www-project-mobile-security-testing-guide/masvs/ (Accessed: 1 September 2025).

Page, M.J., Moher, D. and McKenzie, J.E. (2021) 'Introduction to PRISMA 2020 and Implications for Research Synthesis Methodologists', *Research Synthesis Methods*, 13, pp. 156–163. doi: 10.1002/jrsm.1535.

Palma, F., Realista, N., Serrão, C., Nunes, L., Oliveira, J. and Almeida, A. (2020) 'Automated Security Testing of Android Applications for Secure Mobile Development', *2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. doi: 10.1109/ICSTW50294.2020.00046.

Peruma, A., Huo, T., Araújo, A.C., Imanaka, J. and Kazman, R. (2024) 'Challenges of Novice Developers in Secure Mobile App Development', *Proceedings of the ACM Conference on Software Engineering Education and Training (CSEE&T)*, pp. 101–110.

Rahkema, K. and Pfahl, D. (2022) 'SwiftDependencyChecker: Detecting Vulnerable Dependencies Declared Through CocoaPods, Carthage and Swift PM', *2022 IEEE/ACM 9th International Conference on Mobile Software Engineering and Systems (MobileSoft)*. doi: 10.1145/3524613.3527806.

Salihu, I.-A., Ibrahim, R., Ahmed, B.S., Zamli, K.Z. and Usman, A. (2019) 'AMOGA: A Hybrid UI Model Generation Tool for Android Apps', *Software Testing, Verification & Reliability*, 29(6), e1715.

Statista (2025) *Mobile App Usage Statistics Worldwide*. Available at: https://www.statista.com (Accessed: 1 September 2025).

Steenhock, M., Li, T., Zhang, R. and Choi, H. (2024) 'DeepVulGuard: LLM-Integrated IDE Security Scanning', *arXiv preprint arXiv:2403.12345*. Available at: https://arxiv.org/abs/2403.12345 (Accessed: 1 September 2025).

Zhao, Y., Yi, G., Liu, F., Hui, Z. and Zhao, J. (2022) 'A Framework for Scanning Privacy Information Based on Static Analysis', *2022 IEEE 22nd International Conference on Software Quality, Reliability and Security (QRS)*. doi: 10.1109/QRS57517.2022.00116.

Zhu, J., Li, K., Chen, S., Fan, L., Wang, J. and Xie, X. (2024) 'A Comprehensive Study on Static Application Security Testing (SAST) Tools for Android', *IEEE Transactions on Software Engineering*. doi: 10.1109/TSE.2024.3488041.

SentinelSphere: AI-Driven Cybersecurity Platform Combining Threat Detection with Security Awareness

Nikolaos Tantaroudas[1], Ilias Karachalios[2] and Andrew McCracken[3]
[1]Institute of Communication and Computer Systems, Athens, Greece
[2]National Technical University of Athens, Leof. Alimou, Katechaki, Athens, Greece
[3]DASKALOS-APPS, Péronnas, France

nikolaos.tantaroudas@iccs.gr
ilias_karachalios@mail.ntua.gr
andrew@daskalos-apps.com

Abstract: The growing complexity of cyber threats coupled with the widening cybersecurity knowledge gap presents new challenges in the organisational security domain of the business sector. This paper introduces SentinelSphere, an innovative platform that redesigns cybersecurity defense by integrating advanced threat detection with interactive cybersecurity awareness education, creating a unified approach to building organisational cyber resilience. SentinelSphere employs an Enhanced Deep Neural Network model with specialised feature engineering to significantly reduce false positives while maintaining high detection accuracy across diverse attack vectors. The system includes a Traffic Light System (TLS) to transform complex threat intelligence into intuitive visual indicators, serving simultaneously as an operational tool for security professionals and an educational interface for non-technical users. This paper further presents a Large Language Model that delivers real-time, context-aware cybersecurity guidance and training. This conversational AI agent operates efficiently on standard enterprise hardware, making advanced security education accessible without requiring specialised infrastructure. SentinelSphere is validated using industry-standard datasets, achieving enterprise-grade performance in threat detection with a 94% F1 score and 69.5% reduction in false positives compared to baseline models. The system successfully processed nearly 11 million security events in 30 minutes, demonstrating scalability for enterprise deployment. This work contributes to the cybersecurity field by demonstrating that effective defense requires not just technological sophistication but also systematic enhancement of human security awareness.

Keywords: Cybersecurity awareness, Real-time anomaly detection, Security education, Large language models, Human-centric threat intelligence, Deep neural networks

1. Introduction

The cybersecurity landscape faces an escalating crisis characterised by increasingly complex attack vectors and a critical shortage of skilled security professionals. Traditional Security Information and Event Management (SIEM) systems, while effective at data aggregation, often overwhelm analysts with false positives and lack the contextual intelligence necessary for rapid threat response. Furthermore, the human factor remains the weakest link in cybersecurity, with 82% of data breaches involving human elements according to recent studies (Verizon, 2023).

The ResilMesh project, funded under the EU Horizon Europe programme (Grant Agreement No. 101119681), has established a comprehensive cybersecurity framework designed to enhance cyber resilience for dispersed, heterogeneous cyber systems. Built on the NATS message streaming framework, ResilMesh provides secure event collection and processing capabilities while implementing resilience engineering best practices including redundancy, segmentation, and dynamic positioning (Bernal et al., 2024) This paper presents SentinelSphere, a next-generation cybersecurity platform developed as an extension to the ResilMesh framework that addresses these challenges through an innovative dual approach: combining advanced AI-driven threat detection with integrated security awareness education. Unlike conventional solutions that treat threat detection and security training as separate domains, SentinelSphere creates a synergistic system where every security event becomes an opportunity for organisational learning and resilience building.

The primary contributions of this work are threefold, with explicit novelty articulated for each: (1) An Enhanced Deep Neural Network (DNN) architecture that achieves 94% F1 score while reducing false positives by 69.5% through innovative feature engineering including HTTP-specific anomaly indicators. The novel contribution extends beyond simply adding HTTP features to a DNN. It lies in the systematic identification and validation of ten application-layer indicators that capture attack semantics invisible to network-layer analysis alone. This enables cross-layer threat correlation that existing single-layer intrusion detection approaches cannot achieve, addressing a recognized gap between network-centric detection and application-aware security monitoring. (2) A Traffic Light System (TLS) that transforms complex threat intelligence into intuitive visual indicators, making security understanding accessible across technical expertise levels; The novelty extends beyond threshold-based

alerting to provide a mathematically grounded risk quantification framework that is simultaneously interpretable by non-technical stakeholders and actionable by security analysts, addressing existing communication gaps between security operations and executive decision-making. (3) A Large Language Model-powered conversational agent based on Microsoft's Phi-4 (Microsoft Research, 2024), optimised through Q4_K_M quantisation for deployment on standard enterprise hardware without GPU requirements. The contribution is not simply adding a chatbot to a dashboard, but demonstrating that sophisticated AI-powered security education can be democratized through careful model selection and compression, removing the infrastructure barrier that has limited adoption of intelligent security assistants to well-resourced organisations.

2. Related Work

The cybersecurity domain has identified proactive threat detection and real-time event processing as critical interventions for mitigating the risks posed by advanced cyber threats such as Advanced Persistent Threats (APTs), DDoS attacks, and phishing schemes. It has been identified that the proactive threat detection and real-time event processing as critical interventions for mitigating the risks posed by advanced cyber threats such as Advanced Persistent Threats (APTs), DDoS attacks, and phishing schemes. The convergence of real-time event processing technologies and machine learning algorithms has obtained significant attention, particularly through platforms that integrate solutions such as AWS Kinesis and predictive machine learning models. Herein, we synthesize recent advancements in proactive threat detection methodologies and event processing, highlighting the current state-of-the-art (SOTA) techniques while discussing the potential of specific platforms in addressing existing limitations.

Proactive threat detection effectively employs real-time event processing technologies to provide timely insights into potential cyber threats. Past studies highlight various mechanisms against DDoS attacks, illustrating the necessity for robust real-time defences due to the diverse and evolving attack vectors encountered today (Taneja., 2023). Recent studies further support this by demonstrating how machine learning classifiers can enhance the detection capabilities of DDoS attacks on network logs, thus promoting the need for rapid analysis of event data to identify potential breaches (Musa et al., 2024). Significant attention has also been directed towards APTs, with recent work proposing a methodology that leverages network flow analysis to detect APT attacks by evaluating changes in normal traffic patterns against established baselines (Duong et al., 2020). This proactive approach enables the integration of advanced processing capabilities and predictive analytics to improve response times for critical incidents.

The application of machine learning to cybersecurity has evolved significantly from rule-based systems to sophisticated deep learning architectures. Past studies have demonstrated the effectiveness of deep learning for intrusion detection, achieving great accuracy. Most approached suffered from high false positive rates in production environments. In addition, Vinayakumar et al. (2019) proposed a hybrid deep learning framework combining CNNs and LSTMs, improving detection of zero-day attacks but requiring substantial computational resources. Our work extends these approaches by incorporating application-layer features, particularly HTTP-specific patterns, enabling cross-layer threat correlation. Unlike prior work that analyses network and application layers independently, SentinelSphere's unified 31-feature space enables detection of sophisticated attacks that manifest across protocol boundaries.

The integration of human factors in cybersecurity has gained prominence following high-profile breaches attributed to social engineering. Aldawood and Skinner (2019) emphasised the importance of continuous security awareness training, while Bada et al. (2019) demonstrated that interactive training significantly reduces susceptibility to phishing attacks. Large Language Models have emerged as promising tools for security education. Past studies have explored GPT-based systems for security questionnaire generation, though deployment required significant computational resources (Szabo et al., 2023). Our approach addresses this limitation through Q4_K_M quantisation of Microsoft's Phi-4 model (Abdin et al., 2024), reducing model size from 28GB to 2.5GB while preserving response quality for cybersecurity guidance tasks. This enables deployment on standard 16GB RAM systems without GPU requirements.

The concept of cyber resilience has evolved from traditional security approaches to encompass preparation, response, and recovery capabilities (Araujo et al., 2024). The NIST Cybersecurity Framework 2.0 provides comprehensive guidance for managing cybersecurity risks (NIST, 2024), while the European approach emphasises resilience in critical infrastructure protection (ENISA, 2023). The ResilMesh platform represents a significant advancement in implementing cyber resilience principles through its distributed architecture and AI-based threat awareness capabilities. Bernal et al. (2024) demonstrated edge-based anomaly detection within the ResilMesh framework, highlighting the importance of distributed processing for critical infrastructure

protection. SentinelSphere builds upon this foundation by adding human-centric security awareness capabilities through LLM-chatbot and HTTP-specific patterns to enhance DNN's capabilities of ResilMesh and as a result its Anomaly Detection module.

3. System Architecture and Design

3.1 Overall Architecture

SentinelSphere adopts a microservices architecture integrated with the ResilMesh security framework, enabling scalable, real-time processing of security events. The system comprises four primary layers: Data Ingestion, Processing and Analysis, Intelligence and Education, and Presentation. The architecture of SentinelSphere is a strategic enhancement to the ResilMesh security framework, integrating as an advanced threat analytics layer while preserving the existing architecture's integrity. SentinelSphere operates through direct subscriptions to ResilMesh's NATS messaging infrastructure, consuming both enriched security events ('enriched_events') and anomaly detection alerts ('ad_events'). This non-intrusive approach allows SentinelSphere to function as a parallel consumer without disrupting established workflows.

The Data Ingestion layer utilises Vector for log collection and transformation, processing diverse data sources including network traffic, application logs, and system events. Events flow through NATS message broker, ensuring reliable, high-throughput message delivery to downstream components. This architecture supports processing rates exceeding 5,000 events per second on standard hardware. Vector serves as the primary data collection and transformation layer, ingesting security events from multiple sources and performing real-time transformations including routing, logging, fusion, filtering, augmentation, reduction, and monitoring.

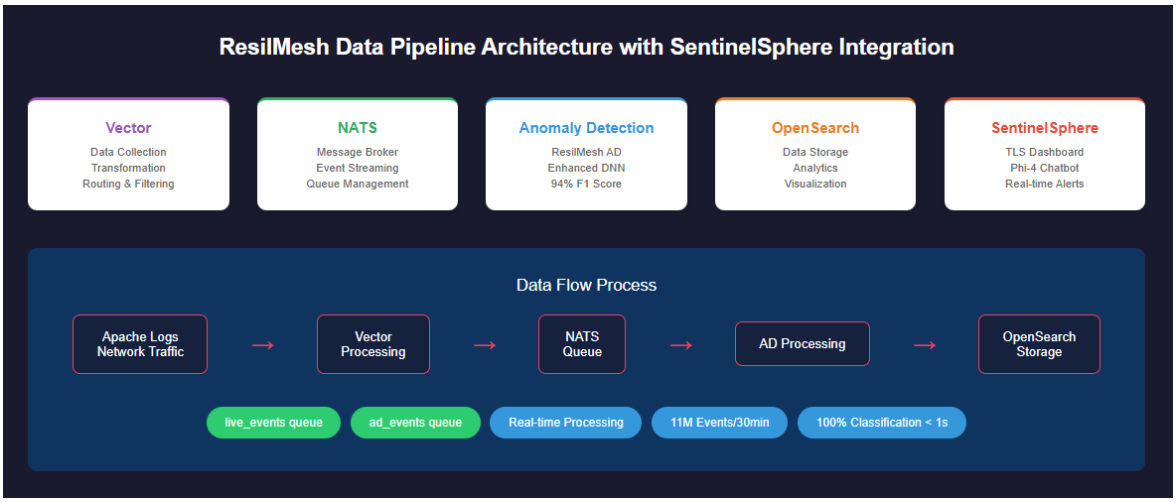


Figure 1: ResilMesh Data Pipeline Architecture - Vector to NATS to OpenSearch Integration

3.2 Enhanced Deep Neural Network Model

The core of SentinelSphere's threat detection capability is an Enhanced DNN model that extends traditional network traffic analysis with application-layer intelligence. The model architecture consists of a 64-64-output neural network enhanced with 31 features: 21 standard network features plus 10 HTTP-specific indicators.

The HTTP feature engineering includes: i) **Request Complexity Score**: Quantifies HTTP request sophistication to identify advanced attacks; ii) **Response Pattern Ratio**: Analyses server response patterns to detect successful exploits; iii) **Path Entropy**: Estimates randomness in request paths, detecting obfuscation attempts; iv) **Automated Tool Detection**: Identifies requests from automated security tools; v) **Attack-Specific Pattern Recognition**: Binary indicators for SQL injection, XSS, and brute force patterns. Training utilised the CIC-IDS2017 and CIC-DDoS2019 datasets (Sharafaldin et al., 2018; Sharafaldin et al., 2019), comprising approximately 400GB of labelled attack data including Web Attack-Brute Force (1,507 samples), Web Attack-XSS (652 samples), Web Attack-SQL Injection (21 samples), and Benign Traffic (168,186 samples).

To address the class imbalance inherent in the dataset, particularly the limited SQL injection samples (n=21), the following methodological safeguards were employed to ensure robustness: i) Stratified splitting: All train/test splits-maintained class proportions using stratified sampling, ensuring minority classes appeared proportionally in both training and evaluation sets. ii) Per-class performance metrics: Beyond aggregate metrics, precision,

recall, and F1 score were computed for each attack class independently (see Table 2), enabling transparent assessment of minority class performance and identification of class-specific weaknesses. iii) Threshold calibration: Rather than using default 0.5 classification thresholds, precision-recall curve analysis was employed on a validation subset to select optimal decision thresholds per class, improving minority class detection. iv) Bootstrap confidence intervals: Performance metrics include 95% confidence intervals computed via 1,000 bootstrap resamples, quantifying uncertainty arising from limited minority class samples.

Model hyperparameters were optimised through grid search on a held-out 20% validation subset: i) Batch size: 32; ii) Learning rate: 0.001 with Adam optimiser; iii) Dropout rate: 0.2; iv) L2 regularisation: 0.001; iv) Maximum epochs: 50 with early stopping (patience=5); v) Decision threshold: 0.4 (calibrated via precision-recall analysis)

3.3 Traffic Light System

The Traffic Light System provides intuitive threat visualisation through a sophisticated scoring algorithm. The system processes anomaly detection events and calculates threat scores from 0-100, determining dashboard status: i) **GREEN (0-30%)**: Normal operation with low threat activity; ii) **YELLOW (30-70%)**: Elevated threat level requiring attention; iii) **RED (70-100%)**: Critical security events demanding immediate response. The TLS scoring algorithm is formally defined as follows:

$$ThreatScore = \min(100, BaseScore \times FrequencyMultiplier \times ClusterFactor \times IP_{Factor} \times DiversityFactor \times SustainedFactor)$$

$$where:$$

$$BaseScore = \frac{\Sigma(anomaly_weight \times confidence \times temporal_weight)}{normalization_factor}$$

Figure 2: Threat Score calculation equation

Table 1: Parameters definition of Traffic Light System of Equation in Figure 2

Parameter	Description	Value Range
BaseScore	Weight sum of detected anomalies normalized to 0-100 scale	0-100
anomaly_weight	Severity weight per threat type (e.g., 0.95 for data exfiltration, 0.90 for unauthorized access, 0.85 for DoS)	0.3-0.95
confidence	Detection confidence score	0.5-1.0
temporal_weight	Time-based decay factor for older events	0-1.0
Normalization_factor	Scaling factor to maintain 0-100 range	Calculated
FrequencyMultiplier	Event rate per minute (1.0x for 1-5 events, 1.5x for 5-20, 2.0x for 20-50, 3.0x for >50)	1.0-3.0x
ClusterFactor	Temporal clustering (1.2x for single cluster, 1.5x for 2 clusters, 2.0x for 3+ clusters)	1.0-2.0x
IPFactor	Attack source concentration (1.5x single IP, 1.3x for 2-4 IPs, 1.0x distributed)	1.0-1.5x
DiversityFactor	Number of different attack types (1.8x for 5+ types, 1.4x for 3-4, 1.2x for 2, 1.0x for 1)	1.0-1.8x
SustainedFactor	Duration of continuous attack activity (2.0x for >3 min high-rate, 1.5x for >2 min, 1.2x for >1 min)	1.0-2.0x

Figure 2 shows the mathematical formula for calculating the final threat score, incorporating base scores, frequency multipliers, clustering factors, IP concentration, diversity factors, and sustained attack indicators. The scoring model employs multi-factor weighted algorithms considering attack type, frequency, clustering, IP concentration, attack diversity, and sustained attack duration. Different anomaly types contribute distinct weights, with data exfiltration and malware receiving the highest weights (0.95) due to critical severity. Threshold Selection Rationale: The 30% and 70% boundaries defining GREEN/YELLOW/RED transitions were determined through empirical analysis during development testing. The 30% threshold was selected to maximise sensitivity to emerging threats while minimising alert fatigue from routine anomalies. The 70% threshold was calibrated to ensure that RED status activations correspond to genuinely critical situations requiring immediate

analyst attention. These thresholds represent operational trade-offs and may be adjusted for specific deployment environments based on organisational risk tolerance.

3.4 LLM-Powered Security Education

SentinelSphere incorporates Microsoft's Phi-4 language model (Abdin et al., 2024), a 14-billion parameter transformer architecture optimised for complex reasoning tasks. Phi-4 was selected based on comparative evaluation against alternative models (Llama-3-8B, Mistral-7B) considering reasoning benchmark performance, parameter count amenable to quantisation, and licensing compatibility for commercial deployment.

The model underwent Q4_K_M quantisation using the llama.cpp framework, reducing model size from 28GB (full precision) to 2.5GB while maintaining acceptable response quality for cybersecurity guidance tasks. Q4_K_M was selected as it provides an optimal balance between compression ratio and quality preservation for instruction-following tasks, with measured perplexity degradation of less than 2%.

Deployment specifications: i) Hardware requirements: Standard business hardware with 16GB RAM, CPU-only (no dedicated GPU required); ii) Inference framework: FastAPI backend integrated with llama.cpp for efficient CPU inference; iii) Performance metrics: 15-20 tokens/second generation rate, time-to-first-token under 2 seconds; iv) Concurrency: Support for 8+ simultaneous conversation streams; v) Memory footprint: 6-8GB during active generation;

The chatbot's cybersecurity knowledge base was constructed from authoritative sources including: i) NIST Cybersecurity Framework 2.0 documentation (NIST, 2024); ii) ENISA threat landscape reports (ENISA, 2023); iii) OWASP Top 10 vulnerability descriptions and remediation guidance; iv) Custom prompt engineering templates for security-contextualised responses.

The chatbot's utility is assessed through its ability to provide contextually appropriate, accurate, and actionable cybersecurity guidance. Qualitative evaluation during development confirmed that the model successfully handles common cybersecurity queries including threat explanation, best practice recommendations, and incident response guidance, while maintaining conversational accessibility for non-technical users.

4. Implementation and Integration

4.1 ResilMesh Integration

SentinelSphere seamlessly integrates with the ResilMesh security framework through subscription to NATS message topics. The system consumes both enriched security events and anomaly detection alerts, enabling comprehensive threat correlation without disrupting existing workflows.

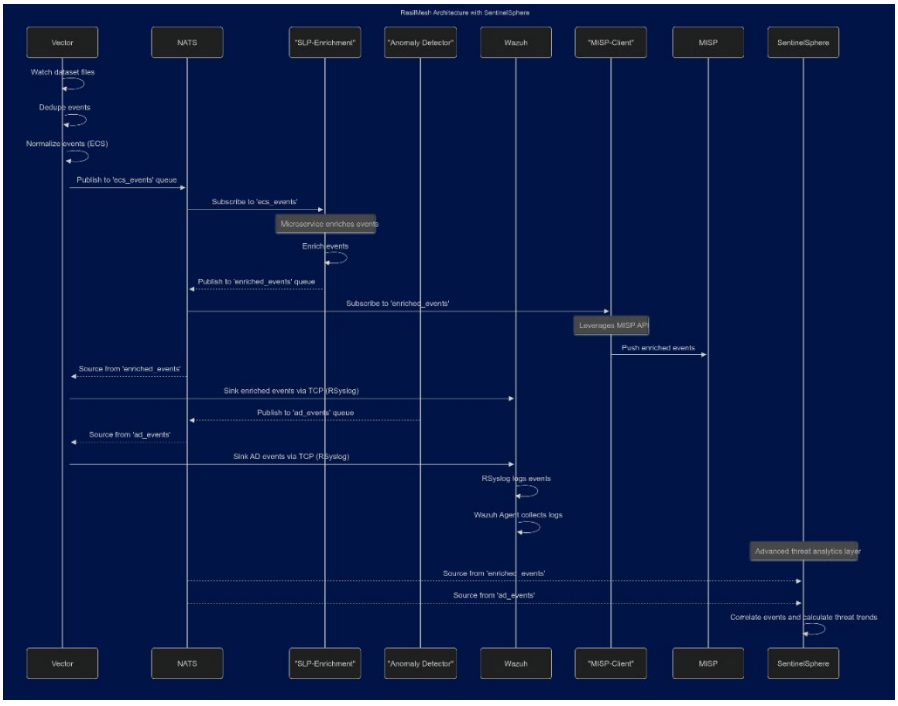


Figure 3: Data Flow Architecture - SentinelSphere and ResilMesh Integration Paths

The diagram in Figure 3 illustrates the comprehensive data flow pipeline where events move from Vector through NATS message broker to the Anomaly Detector, then through ad_events topic to the Threat Calculator, which updates Redis before displaying on the Dashboard. Docker containerisation ensures deployment consistency across environments. The implementation includes environment variable configuration, network isolation, and resource management aligned with ResilMesh standards. This architectural alignment facilitates modular deployment within broader security ecosystems.

4.2 Data Pipeline Implementation

The data pipeline leverages AWS OpenSearch for scalable indexing and analysis. Vector serves as the primary collection layer, performing real-time transformations including routing, filtering, augmentation, and monitoring. Events flow through NATS to the Anomaly Detection module, with processed events stored in OpenSearch for historical analysis. Performance optimisation through grid search yielded key parameters: i) Base threshold: 0.3 for normal conditions; ii) Warning threshold: 0.6 for elevated risk; iii) Critical threshold: 0.8 for high-risk conditions; iv) Burst detection window: 5-minute intervals; v) Minimum event count: 10 events for status changes.

4.3 Dashboard Implementation

The SentinelSphere dashboard provides comprehensive security visualisation and interaction capabilities through a web-based interface built with FastAPI and vanilla JavaScript. The dashboard displays real-time threat level indicators using the Traffic Light System, along with historical trend analysis showing detected threats and vulnerabilities based on real-time data.

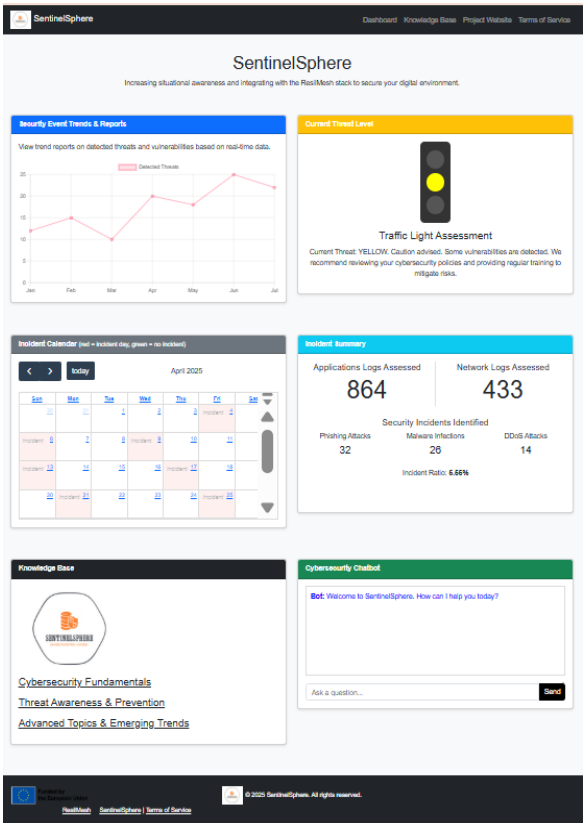


Figure 4: SentinelSphere Dashboard-HomePage and chatbot interaction panel

Figure 4 shows the incident calendar for temporal analysis, the security incident summary statistics, and the integrated Phi-4 powered cybersecurity chatbot providing real-time assistance. The dashboard implements key functionalities including real-time event viewing, TLS indicator panels displaying dynamic threat levels, forecast and trend charts for security pattern visualisation, and an interactive chatbot panel for AI assistance and user query handling.

5. Experimental Evaluation

5.1 Dataset and Methodology

Evaluation utilised the CIC-IDS2017 dataset containing 170,366 network flows with labelled cyberattacks. The benchmark methodology involved training parameter optimisation on 20% of logs, with performance evaluation on the remaining 80% to minimise overfitting. Both baseline and enhanced models used identical 64-64-output DNN architecture, differing only in feature sets.

5.2 Performance Results

Table 2 presents comparative performance metrics between the baseline ResilMesh model and SentinelSphere's enhanced model, including 95% bootstrap confidence intervals:

Table 2: Performance Comparison

Metric	Baseline DNN	Enhanced DNN	Improvement
F1 Score	91.1% ± 0.8%	94.0%± 0.6%	3.3%
Accuracy	96.4%± 0.4%	97.7%± 0.3%	1.3%
Precision	91.0%± 1.1%	96.9%± 0.7%	6.5%
Recall	91.1%± 0.9%	91.3%± 0.8%	0.2%
False Positive Rate	2.3%	0.7%	69.5%
Brute Force Detection	90.3%	91.2%	0.9%
XSS Detection	91.8%	94.4%	2.6%
SQL Injection Detection	85.7%	90.5%	4.8%

Table 3: Per-Class Performance Metrics (Enhanced DNN Model)

Attack Class	Samples (n)	Precision	Recall	F1 Score	95% CI (F1)
Benign	168,186	98.2%	99.1%	98.6%	± 0.2%
Brute Force	1,507	92.4%	91.2%	91.8%	± 1.8%
XSS	652	95.1%	94.4%	94.7%	± 2.3%
SQL Injection	21	88.9%	90.5%	89.7%	± 8.4%

The wider confidence interval for SQL Injection (±8.4%) reflects the limited sample size (n=21) in the dataset. Despite this methodological constraint, the enhanced model achieves 90.5% detection rate, representing a 4.8 percentage point improvement over the baseline (85.7%). This improvement is attributable to the HTTP-specific features that capture SQL injection patterns including parameterised query anomalies and response pattern changes that are invisible to network-layer-only analysis. The enhanced model demonstrates significant improvement in precision and false positive reduction while maintaining high recall rates. Most notably, false positives decreased from 59 to 19 instances, representing a 69.5% reduction critical for SOC operational efficiency.

5.3 Scalability Testing

Comprehensive load testing validated enterprise enterprise scalability using the following hardware and deployment configurations: Hardware Specifications: i) CPU: Intel Core i7-10700 (8 cores, 16 threads, 2.9GHz base clock, 4.8GHz turbo); ii) RAM: 32GB DDR4-3200; iii) Storage: 512GB NVMe SSD (Samsung 970 EVO Plus, sequential read 3,500 MB/s); iv) Operating System: Ubuntu 22.04 LTS; v) Container Runtime: Docker 24.0.5 with Docker Compose orchestration. Deployment Configuration: i) 5 Docker containers (Dashboard, Chatbot, Vector, NATS, OpenSearch); ii) NATS message broker configured with 4 worker threads; iii) OpenSearch single-node cluster with 2GB heap allocation; iv) Redis 7.0 for TLS score caching; v) 8 parallel event processing streams

scalability. The system successfully processed 10,998,214 events in approximately 30 minutes, representing three months of Apache log data. This translates to sustained processing exceeding 5,000 events per second on standard desktop hardware.

Table 4: Docker container resource consumption under load

Component	CPU Usage	Memory	Startup Time
Dashboard	15%	512MB	3 seconds
Phi-4 Chatbot	40%	2.5 GB	8 seconds
Vector	20%	256MB	2 seconds
NATS	10%	128MB	1 second
OpenSearch	25%	2GB	10 seconds

5.4 Chatbot Performance Evaluation

The Phi-4 powered cybersecurity chatbot demonstrated effective domain-specific responses with practical deployment metrics.

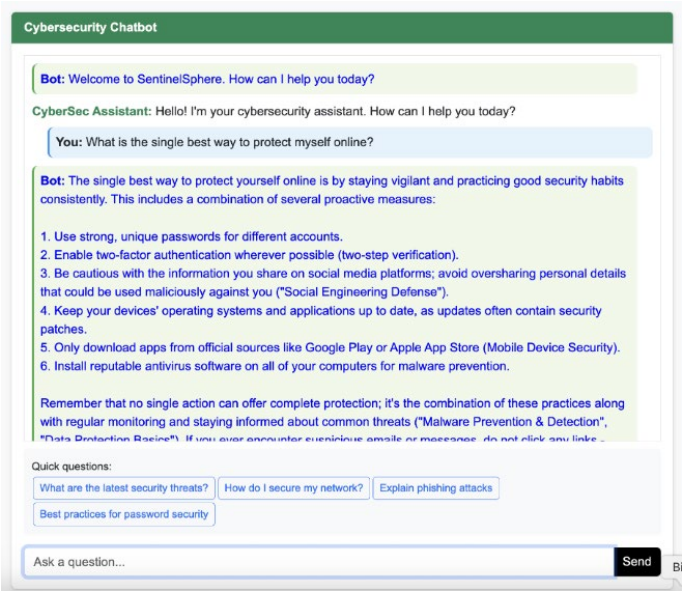


Figure 5: Demonstration of Phi-4 Cybersecurity domain specific LLM Response

The chatbot in Figure 5 provides comprehensive cybersecurity guidance, explaining best practices for online protection including strong passwords, two-factor authentication, cautious information sharing, system updates, app downloads from official sources, and antivirus software installation. The chatbot successfully handles complex cybersecurity queries while maintaining conversational flow and providing actionable recommendations tailored to user expertise levels.

6. Discussion and Implications

6.1 Addressing the Human Factor

SentinelSphere's integration of threat detection with security education represents a paradigm shift in cybersecurity defense. Traditional approaches treat security awareness as separate from operational security, creating disconnects between threat detection and human response. Our unified approach ensures every security event contributes to organisational learning, progressively building cyber resilience.

The LLM-powered chatbot broadens accessibility to security expertise, enabling non-technical users to understand and respond to threats effectively. Similar approaches have demonstrated positive impacts in other domains, where targeted, technology-enhanced education successfully raised awareness and improved response behaviors (Karachalios, 2024; Karachalios & Tantaroudas, 2025). By operating on standard hardware, the solution remains accessible to organisations without specialised infrastructure, addressing the cybersecurity skills gap plaguing many enterprises.

6.2 Practical Deployment Considerations

The 69.5% reduction in false positives has profound implications for SOC operations. Security analysts spend significant time investigating false alerts, leading to alert fatigue and missed genuine threats. SentinelSphere's enhanced precision enables analysts to focus on legitimate threats, improving both efficiency and effectiveness. The Traffic Light System's intuitive visualisation facilitates rapid threat assessment across organisational hierarchies. Executive stakeholders can understand security posture without technical expertise, enabling informed decision-making during critical incidents.

6.3 Integration with ResilMesh Ecosystem

SentinelSphere's seamless integration with the ResilMesh framework demonstrates the value of modular cybersecurity architectures. By leveraging ResilMesh's NATS-based messaging infrastructure and resilience engineering principles, SentinelSphere adds advanced threat analytics and human-centric capabilities without disrupting existing security workflows. This integration approach aligns with recent cybersecurity mesh concepts (Ramos-Cruz et al., 2024), enabling distributed security services while maintaining centralised intelligence and coordination. The success of this integration validates the ResilMesh architecture's extensibility and its potential for accommodating diverse security solutions.

6.4 Limitations and Future Work

While SentinelSphere demonstrates significant advances, several areas warrant future investigation. The current implementation focuses on HTTP-based attacks; extension to other protocols would enhance coverage. Additionally, the chatbot's knowledge base, while comprehensive for common threats, requires continuous updating for emerging attack vectors. The limited SQL injection samples (n=21) in the CIC-IDS2017 dataset represents a methodological constraint. While addressed through stratified splitting and per-class metrics with confidence intervals, future work should incorporate additional datasets with more balanced attack class distributions to strengthen validation of minority class detection performance. Future work will explore federated learning approaches, enabling organisations to benefit from collective threat intelligence while maintaining data privacy. Integration with automated response systems could further reduce response times for known attack patterns.

7. Conclusion

This paper presented SentinelSphere, an innovative cybersecurity platform that successfully bridges the gap between advanced threat detection and human security awareness. Through careful integration with the ResilMesh framework and implementation of Enhanced Deep Neural Networks, intuitive threat visualisation, and AI-powered security education, the system addresses critical challenges in modern cybersecurity defence. Key achievements include a 94% F1 score in threat detection with 69.5% reduction in false positives, successful processing of nearly 11 million events in 30 minutes, and deployment of sophisticated AI capabilities on standard enterprise hardware. These results demonstrate that effective cybersecurity requires not just technological sophistication but systematic enhancement of human security awareness.

SentinelSphere's dual approach—treating every security event as both a threat to mitigate and an opportunity to educate—represents a fundamental shift in cybersecurity philosophy. By democratising security understanding and making advanced threat intelligence accessible across expertise levels, the platform contributes to building more resilient organisations capable of defending against evolving cyber threats. The successful integration with ResilMesh validates the importance of modular, extensible cybersecurity architectures that can accommodate innovative solutions while maintaining operational stability. As cyber threats continue to evolve, platforms like SentinelSphere that combine technological advancement with human empowerment will be essential for maintaining effective cyber defence. This assertion is widely supported in current literature, with recent research emphasizing that integrated, AI-driven platforms and cybersecurity mesh architectures are fundamental to effective cyber defense in contemporary organizational environments (Jada & Mayayise, 2024).

Acknowledgements

This work was supported by the European Union's Horizon Europe research and innovation programme under the ResilMesh project (Grant Agreement No. 101119681).

Ethics Declaration: This research did not require ethical clearance as all experiments were conducted using publicly available datasets with no human participants involved in data collection.

References

Abdin, M., Jacobs, S.A., Amin, A.A., Aneja, J., Awadalla, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Beber, J. .(2024)"Phi-4 Technical Report", arXiv preprint arXiv:2412.08905. https://doi.org/10.48550/arXiv.2412.08905

Aldawood, H. and Skinner, G. (2019) "Reviewing Cyber Security Social Engineering Training and Awareness Programs—Pitfalls and Ongoing Issues", *Future Internet*, Vol 11, No. 3, pp 73-89. https://doi.org/10.3390/fi11030073

Araujo, M. S. d., Machado, B. A. S., & Passos, F. U. (2024). Resilience in the Context of Cyber Security: A Review of the Fundamental Concepts and Relevance. Applied Sciences, 14(5), 2116. https://doi.org/10.3390/app14052116

Bada, M., Sasse, A.M. and Nurse, J.R. (2019) "Cyber Security Awareness Campaigns: Why do they fail to change behaviour*?", International Conference on Cyber Security for Sustainable Society*, pp 118-131. https://doi.org/10.48550/arXiv.1901.02672

Bernal, J. Fernandez, P., Montoro, D., Lee, B., Lan, Xi., Stojanovic, B., Jorgeley, Husak., M., ResilMesh D2.2 System Architecture (2024) "ResilMesh: Situation Aware enabled Cyber Resilience for Dispersed, Heterogenous Cyber Systems", EU Horizon Europe Project 101119681, Public Deliverable.

ENISA (2023) "ENISA Threat Landscape 2023", European Union Agency for Cybersecurity, Luxembourg.

Friedberg, I., Skopik, F., Settanni, G., & Fiedler, R. (2015). Combating advanced persistent threats: from network event correlation to incident detection*. Computers & Security*, 48, 35-57. https://doi.org/10.1016/j.cose.2014.09.006

https://aws.amazon.com/opensearch-service/

Jada, I., & Mayayise, T. O. (2024). The impact of artificial intelligence on organisational cyber security: An outcome of a systematic literature review. *Data and Information Management*, *8*(2), 100063. https://doi.org/10.1016/j.dim.2023.100063

Karachalios, I. (2024). Utilizing Educational Gaming to Foster Sustainability Awareness in Corporate Settings. *International Journal of Science and Research (IJSR)*, 13(3), 740–744. https://doi.org/10.21275/SR24308032254

Karachalios, I., & Tantaroudas, N. (2025). Future Greek Pre-Service Teachers' Knowledge, Attitudes and Self-Efficacy in Waste Management. *British Journal of Education,* 13(8), 25–42. https://doi.org/10.37745/bje.2013/vol13n82542

Kaur, P., Kumar, M., & Bhandari, A. (2017). A review of detection approaches for distributed denial of service attacks. *Systems Science & Control Engineering*, 5(1), 301-320. https://doi.org/10.1080/21642583.2017.1331768

Musa, M. and Odokuma, E. (2024). A framework for the detection of distributed denial of service attacks on network logs using ml and dl classifiers. *Scientia Africana*, 22(3), 153-164. https://doi.org/10.4314/sa.v22i3.14 Duong, L. (2020). Detecting APT attacks based on Network Flow. *International Journal of Emerging Trends in Engineering Research*, 8(7), 3134-3139. https://doi.org/10.30534/ijeter/2020/42872020

NIST (2024) "NIST Cybersecurity Framework 2.0", *National Institute of Standards and Technology, U.S. Department of Commerce, Washington DC*.

Ramos-Cruz, B., Martínez, L. and Wang, H. (2024) "The Cybersecurity Mesh: A Comprehensive Survey of Involved Artificial Intelligence Methods", *Computer Networks*, Vol 223, pp 109-124. https://doi.org/10.1016/j.neucom.2024.127427

Riskhan, B., Safuan, H., Hussain, K., Elnour, A., Abdelmaboud, A., Khan, F., … & Kundi, M. (2023). An adaptive distributed denial of service attack prevention technique in a distributed environment. *Sensors*, 23(14), 6574. https://doi.org/10.3390/s23146574

Sadlek, L., Husák, M. and Čeleda, P. (2024) "Hierarchical Modeling of Cyber Assets in Kill Chain Attack Graphs", 2024 *20th International Conference on Network and Service Management (CNSM)*, IEEE, pp 1-9. 10.23919/CNSM62983.2024.10814501

Sharafaldin, I., Lashkari, A.H. and Ghorbani, A.A. (2018) "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, pp. 108-116. https://doi.org/10.5220/0006639801080116*

Sharafaldin, I., Lashkari, A.H., Hakak, S. and Ghorbani, A.A. (2019) "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy", IEEE 53rd International Carnahan Conference on Security Technology*, pp. 1-8. https://doi.org/10.1109/CCST.2019.8888419*

Sinha, S. and Degadwala, S. (2024). Voting model strategies for reliable categorical iot-ddos attack prediction. *International Journal of Scientific Research in Computer Science Engineering and Information Technology*, 10(2), 300-307. https://doi.org/10.32628/cseit2410223

Szabó, Z., & Bilicki, V. (2023). A New Approach to Web Application Security: Utilizing GPT Language Models for Source Code Inspection. Future Internet, 15(10), 326. https://doi.org/10.3390/fi15100326

Taneja, N. (2023). A deep dive into methods for combating ddos attacks and securing data. International Journal of Computing Programming and Database Management, 4(2), 01-07. https://doi.org/10.33545/27076636.2023.v4.i2a.84

Verizon (2023) "2023 Data Breach Investigations Report", Verizon Business, New York.

Vinayakumar, R., Alazab, M., Soman, K.P., Poornachandran, P., Al-Nemrat, A. and Venkatraman, S. (2019) "Deep Learning Approach for Intelligent Intrusion Detection System", *IEEE Access*, Vol 7, pp 41525-41550. 10.1109/ACCESS.2019.2895334

Yin, C., Zhu, Y., Fei, J. and He, X. (2017) "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks", *IEEE Access*, Vol 5, pp 21954-21961. 10.1109/ACCESS.2017.2762418

Aligning DYNAMO Framework with the EU Cyber Resilience Act in the Energy Sector

Ilkka Tikanmäki[1,2][1] Jarmo Maikkola[1], Joonas Nykopp[1], Petri Nyfors[1], Sara Väisänen[1] and Shakti Panta Khatri[1]

[1]Security and Risk Management, Laurea University of Applied Sciences, Espoo, Finland
[2]Department of Warfare, National Defence University, Helsinki, Finland

Ilkka.tikanmaki@laurea.fi

Abstract: The importance of robust cybersecurity frameworks has been raised by the digitalisation of critical infrastructure, particularly in the energy sector. The European Union (EU) launched the Cyber Resilience Act (CRA) in 2022, establishing uniform cybersecurity standards for products with digital elements at all lifecycle stages to address this issue. CRA describes requirements for software and hardware products with digital elements placed on the EU market. This study examines the CRA's effects on the energy sector and evaluates how the DYNAMO platform can support compliance and enhance sectoral resilience. The platform's key element is a dynamic resilience assessment methodology, which combines business continuity management (BCM) and cyber threat intelligence (CTI). Significant cybersecurity vulnerabilities in the energy sector are identified in the study, which include a growing attack surface, complex supply chains, and convergence of operational technology (OT) and information technology (IT) systems. CRA's inability to address OT-specific challenges, particularly in legacy systems like SCADA, is highlighted in the study through a literature review and case study analysis. The gap analysis shows that although CRA follows standards like NIST and ISO 27001, it doesn't have provisions for real-time monitoring, adaptive risk management, and OT-specific protections. To resolve those gaps, the research suggests that DYNAMO include 24-hour incident reporting to the European Union Agency for Cybersecurity (ENISA), structured vulnerability disclosure protocols, and post-market surveillance mechanisms. Additionally, DYNAMO must develop customised plans for OT environments, which involve retrofitting outdated systems and improving threat detection abilities. The findings show that cybersecurity in the energy industry requires a more dynamic and functionally integrated approach. Aligning DYNAMO and CRA will support regulatory compliance and strengthen the industry's resilience to evolving cyber threats. The next stage of research should be to validate these recommendations via empirical testing and explore cross-sector applications of the DYNAMO framework.

Keywords: Cyber resilience act (CRA), Energy sector cybersecurity, Operational technology (OT), Critical infrastructure protection, Dynamo platform, post-market monitoring

1. Introduction

1.1 Background and Research Problem

The digitalisation of the critical infrastructure, especially the energy sector, has increased reliance on digital systems. The shift to more digital systems has exposed vulnerabilities in the energy sector and other critical infrastructure. This has made cybersecurity resilience an important part of comprehensive security. The European Union introduced the Cyber Resilience Act (CRA) in 2022 to establish uniform cybersecurity standards. CRA is a legal framework that describes requirements for software and hardware products with digital elements placed on the European Union market. These requirements cover the product lifecycle. (European Commission. Joint Research Centre, 2024)

The energy sector's strategic importance and its heightened vulnerability to cyber threats and regulatory complexities, examining how frameworks like the Dynamic Resilience Assessment Method Including Combined Business Continuity Management and Cyber Threat Intelligence solution for Critical Sectors (DYNAMO) project, align with CRA compliance, has become important so that compatibility with DYNAMO and CRA can be verified (DYNAMO project, 2024). DYNAMO is a dynamic resilience assessment method. It combines business continuity management (BCM) and cyber threat intelligence (CTI); these features can create a situational awareness picture and enhance organisational resilience. (DYNAMO project, 2024).

This study contributes to the growing body of knowledge on cybersecurity regulation and resilience of critical infrastructure by providing a comparative analysis of the EU Cybersecurity Act and the DYNAMO framework. It highlights regulatory gaps in the CRA, its limited coverage of operational technology, and suggests improvements to DYNAMO that would enable better alignment with CRA requirements. By combining business continuity management and cyber threat intelligence, the study presents a dynamic approach to resilience that

[1]https://orcid.org/0000- 0001-8950-5221

complements the static compliance mechanisms of the CRA. This dual perspective contributes to understanding how regulatory and operational frameworks can be aligned to improve cybersecurity in the energy sector.

Although CRA outlines mandatory cybersecurity measures across sectors (Szedlak, Reinemann and Hatzelmann, 2024), existing literature seems to lack a detailed analysis of how sector-specific frameworks, such as DYNAMO, align with CRA requirements - especially regarding unique challenges faced by the energy sector. This study addresses this gap by exploring CRA's practical implications for energy cybersecurity and evaluating how DYNAMO can effectively support regulatory compliance and resilience improvements.

Research questions of this study are:

RQ1: How does the Cyber Resilience Act (CRA) impact cybersecurity policies and resilience in the energy sector?

RQ2: How can existing frameworks like DYNAMO support compliance?

This study hypothesises that the CRA introduces stricter compliance measures expected to improve cybersecurity resilience in the energy sector, but implementation challenges may limit its full effectiveness. This study examines the requirements of the EU Cyber Resilience Act from the perspective of cybersecurity standards and cybersecurity risks.

This study addresses several pressing challenges in aligning cybersecurity frameworks with regulatory requirements in the energy sector. In the EU CRA, lifecycle-based cybersecurity obligations are primarily designed for IT systems, but they are not sufficiently covered for operational technology (OT), particularly legacy systems like SCADA. Real-time monitoring and adaptive response mechanisms are necessary in the energy sector due to the rapidly evolving threat landscape, which includes ransomware, supply chain attacks, and AI-based attacks. The formal compliance mechanisms and post-market monitoring capabilities required by the CRA are not present in existing frameworks like DYNAMO, even though they are strong in terms of resilience and situational awareness. By examining how DYNAMO can be improved to meet the CRA requirements, the study aims to improve regulatory compliance and operational resilience in critical energy infrastructures.

After the introduction, Section 2, a literature review, reviews the current state of the cyber environment and vulnerabilities in the energy sector, comparing the differences between DYNAMO and CRA using GAP analysis. Section 3 describes the methodology used. Section 4, Results, presents the research findings on cybersecurity threats in the energy sector and gaps concerning modern cyber threats. Finally, Section 5 summarises the results and provides recommendations for future research.

2. Literature

In the cyber threat landscape, the energy sector forms a high-risk target. The backbone of modern society is based on electricity, which sustains industry, households, and critical infrastructure such as healthcare and transport. In the energy sector, production, storage, and distribution are increasingly based on network- and software-based solutions, which is why they are exposed to vulnerabilities (ECSO, 2018). New technology also brings new types of threats, such as attack methods based on artificial intelligence. Also, the geopolitically unstable situation in the world increases the risk of large-scale and disruptive cyberattacks targeting the energy sector (World Economic Forum, 2025). Especially in the electricity sector, the resilience of the electricity ecosystem has been found, according to research, to be weak, and in addition, there is a shortage of cyber professionals on the market (Ruoslahti and Tikanmäki, 2022).

Cyberattacks have various consequences, such as impacts on the economy, national security, daily functions, and overall trust in society. Economically, attacks can cause significant damage, increase energy costs, and disrupt trade. An attack on energy infrastructure affects national defence and security policies. When household access to electricity becomes more difficult, daily services become more complicated, endangering citizens' safety and reducing the quality of life. In the long run, the unreliability of the energy system and frequent power outages erode trust among businesses and citizens in society’s stability.

CRA obligations include meeting security requirements relating to the properties of products with digital elements, such as ensuring the absence of any known exploitable vulnerabilities, providing secure-by-default configurations, and enabling the ability to address vulnerabilities through security updates. Additionally, they must comply with vulnerability handling requirements, which involve conducting regular testing and security reviews of the product. Manufacturers are required to manage and address vulnerabilities effectively for the expected lifetime of the product, or five years from the product’s placement on the market, whichever is shorter. (Schmitz-Berndt and Cole, 2022).

Ukraine's electricity grid was subjected to a cyberattack on 23 December 2015; the attack targeted three regional electricity companies in Ukraine. As a result of the attack, 225,000 customers were left without electricity for several hours. The attack took place at three different sites almost simultaneously, indicating a precisely planned and coordinated operation. The implementation of the attack included, among other things, a telephone denial-of-service attack, malicious firmware, the KillDisk destruction program, and the BlackEnergy 3 malware (Lee, Assante and Conway, 2016).

In the United States, the Colonial Pipeline company was subjected to a ransomware attack on May 7, 2021; behind the attack was the DarkSide group. The attack targeted the company's billing system. Approximately 45% of the gasoline and jet fuel on the U.S. East Coast flowed through the company's pipeline. The entire pipeline had to be shut down, which affected the lives of millions of people, airports, and airlines, resulting in increased gas prices, fuel shortages at gas stations, and panic. The company paid the group a ransom of 4.4 million dollars, part of which was recovered, but the company's brand suffered significant reputational damage. The attack's success was enabled by inadequate protection of multi-factor authentication, a basic cybersecurity tool against hackers (Ene and Savu, 2023). The CRA explicitly introduces a product lifecycle approach, recognising that technological products can evolve. They may become insecure or be used in new contexts, requiring continuous attention to cybersecurity.

3. Method

This study utilises a descriptive literature review and case study analysis. The descriptive literature review systematically examines existing research and reports (Kim, Sefcik and Bradway, 2017) related to the Cyber Resilience Act, cybersecurity frameworks, and threats in the energy sector. The case study method (Yin, 2009) provides detailed examinations of significant cybersecurity incidents, enabling an understanding of real-world impacts, vulnerabilities, and the practical implications of CRA compliance and implementation within critical energy infrastructures.

This study employs a qualitative research design involving a descriptive literature review and case study analysis to examine the connection between the Cyber Resilience Act (CRA) and the DYNAMO framework for energy cybersecurity. A systematic review of literature involves examining academic publications, EU regulatory documents, cybersecurity frameworks such as NIST (National Institute of Standards and Technology, 2018), ISO 27001 (ISO, 2022), and industry reports (Saeed *et al.*, 2023). Sources were selected based on their relevance to CRA implementation, energy sector vulnerabilities, and resilience strategies. The evaluation identifies gaps in CRA's operational technology (OT) coverage and assesses DYNAMO's response to these challenges.

Real-world vulnerabilities and the consequences of inadequate cybersecurity measures were illustrated through the analysis of two high-profile cyberattacks, the 2015 Ukraine power grid attack and the 2021 Colonial Pipeline ransomware attack. To mitigate similar threats, these cases are used to evaluate the practical applicability of CRA and the potential of DYNAMO. A gap analysis was done to compare CRA requirements and DYNAMO capabilities. The objective was to compare CRA's lifecycle security obligations to DYNAMO's BCM and CTI capabilities, and identify any overlaps, gaps, or potential improvements. The analysis concentrated on reporting incidents, post-market monitoring, and protections specific to the purpose.

4. Results

4.1 Cybersecurity Threats in the Energy Sector

In the energy sector, three essential characteristics expose companies to cyberattacks. The attack surface has increased because companies' operations and infrastructure have been spread over a wide geographical area. The actors in the field and their relationships with third-party supply chains are complex, exposing supply chains to supply chain attacks. The physical and digital infrastructures in the electricity and gas sector are interconnected, which results in the OT infrastructure and IT networks being exposed to attacks.(Ene and Savu, 2023). Table 1 describes types of attacks that target the energy sector.

Table 1. Types of attacks targeting the energy sector (Ene and Savu, 2023).

Type of attack	Description
Malware	Software affecting the computer
Web-based attacks	Website used as a tool
Social engineering	Psychological manipulation

Type of attack	Description
Denial-of-service attacks (DoS)	Disable the user's computer or network
Insider threats	Parties connected to the organisation's network
Cyber espionage and cyber warfare	Unauthorised user activity
Botnets	Devices infected with malware
Ransomware	Malware is preventing the victim from accessing data

Malware is any software designed to affect a computer and install itself without the user's permission. Web-based attacks are threats in which a website is used as a tool of cybercrime. Social engineering (e.g., phishing, spam) is a psychological manipulation in which the attacker poses as a trustworthy party to get the target to perform desired actions, such as revealing confidential information or granting access to an account and computer. Denial-of-service attacks (DoS) aim to temporarily or permanently disable the user's machine or network by making it unusable. Insider threats (e.g., careless employees) include all parties connected to the organisation who may have critical information about the organisation's security practices, data, and IT systems. Cyber espionage and cyber warfare are attacks that involve unauthorised user activity in which data or systems are stolen from individuals or organisations for financial, personal, political, or military gain. Botnets, in which all devices connected to the network are infected with malware, enable remote control of the devices. Ransomware means malware that prevents the victim from accessing critical data or systems by encrypting them until the victim pays a ransom (Ene and Savu, 2023).

4.2 Overview of the Cyber Resilience Act (CRA)

The CRA (Cyber Resilience Act) applies to products with digital elements, including any software or hardware product and its associated remote data processing solutions. Several obligations must be fulfilled both before and while placing a product with digital elements on the market. For example, manufacturers, importers, and distributors must ensure that the product is accompanied by appropriate instructions and information, provided in a language that is easy for users to understand, to guarantee safe usage. (Schmitz-Berndt and Cole, 2022).

Manufacturers should carry out a cybersecurity risk assessment for each product, considering the assessment throughout all stages of the product's lifecycle - from development to delivery and maintenance. This risk assessment plays a crucial role in helping manufacturers determine the "appropriate level of cybersecurity based on the risks" (European Parliament and of the Council, 2024, p. 68) and identify the specific cybersecurity measures, as prescribed by the CRA, that must be implemented for the product. By adopting a risk-based approach, the CRA promotes a regulatory framework that adapts requirements according to the level of risk, while ensuring a consistent and coherent method for regulating the underlying technology. (Shaffique, 2024).

As regards importers, the Act provides for the obligation to ensure compliance with the essential cybersecurity requirements set out in Annexe I of the Act. Importers must check that the manufacturer has carried out a conformity assessment, confirm that the technical documentation is available, and that the product bears the Conformité Européenne (CE) marking, provide the necessary contact details, and include user-friendly instructions and product information. (Kaminaris, 2025).

To ensure that users can respond promptly to security incidents, manufacturers should inform users about any such incident and, where applicable, provide details of corrective measures to mitigate its impact. This communication can be done, for example, by publishing relevant information on the manufacturer's website, or where the manufacturer can contact users and justified by the risks by reaching out to users directly. (Schmitz-Berndt and Cole, 2022).

4.3 CRA's Limitations

The CRA framework presents an organised approach to evaluate cybersecurity readiness for the energy sector, which, however, has considerable gaps in contemporary cyber threats. The periodic nature of the audit is one major limitation; real-time monitoring is what the energy infrastructure needs against threats that evolve quickly, such as ransom attacks and supply chain compromises. While the CRA does fit well with NIST and ISO 27001 compliance standards, they often focus on regulatory checkbox exercises that detract from real security effectiveness, thus creating blind spots against advanced persistent threats (National Institute of Standards and Technology, 2018; ISO, 2022). Above all, it is rather more designed from an IT perspective, neglecting unique vulnerabilities arising from operational technology (OT) environments, particularly legacy ICS/SCADA systems, which are very much in use in energy networks (Stouffer *et al.*, 2023). These deficiencies highlight the

importance of an ever-evolving CRA that exceeds the behavioural static measures of compliance and meets the dynamic state of the challenge for energy sector cybersecurity.

To improve CRA's efficacy, the energy sector shall integrate within the framework capabilities for continuous threat monitoring and adaptive risk management. This means building security modules specific to OT, establishing dynamic threat intelligence feeds, and conducting mandatory cyber war-gaming exercises to stress-test incident response plans. By closing these gaps, CRA can evolve from a compliance tool to an active, real-time cybersecurity operational tool for protecting critical energy infrastructure. The modernisation of the CRA alongside investigators, energy companies, and information security personnel would have to occur to keep pace with the fast-changing threats facing the energy sector. Improvements would enable CRA to assess not just resilience but also actively reinforce defences against new emerging cyber threats.

4.4 Key Implications of CRA and DYNAMO

The energy sector faces several strategic and operational implications due to the comparison between the CRA and the DYNAMO framework. Table 2 summarises the key implications of the CRA and DYNAMO frameworks, highlighting their respective strengths and limitations.

Table 2*:* Key implications of CRA and DYNAMO.

Aspect	Cyber Resilience Act (CRA)	DYNAMO Framework
Purpose	Regulatory compliance for products with digital elements	Operational resilience for critical infrastructure
Focus Area	IT systems and product lifecycle	IT + OT systems, situational awareness, and threat intelligence
Approach	Static, audit-based, lifecycle security	Dynamic, real-time monitoring and adaptive response
OT Coverage	Limited; primarily IT-centric	Designed to include OT, but needs enhancement for legacy systems
Incident Reporting	Mandatory 24-hour reporting to ENISA	Not currently integrated; recommended for future alignment
Post-Market Monitoring	Required throughout product lifecycle	Not built-in; suggested for integration
Compliance Mechanism	CE marking, conformity assessments, documentation	No formal compliance mechanism; focuses on resilience
Threat Intelligence	Not included	Core component (CTI)
Lifecycle Security	Mandatory	Not explicitly covered

The energy sector faces multiple challenges in cybersecurity strategy, regulatory compliance, and operational resilience due to the differences between the DYNAMO framework and the CRA. Legal accountability is ensured by the CRA's compliance-based approach, but real-time protection against evolving threats is not guaranteed. DYNAMO's design is focused on resilience and provides proactive defence, but does not have formal regulatory compliance. Energy companies can fulfil the CRA's requirements without being completely secure, or they can be resilient but not compliant, and both situations have risks. The CRA's OT system vulnerabilities that are common in energy infrastructure are ignored by its IT-centric focus. DYNAMO's ability to tackle OT threats is crucial but not yet fully developed.

Legacy systems are vulnerable without OT-specific measures, which increases the risk of attacks like those in Ukraine and the Colonial Pipeline. Rapidly evolving threats may prevent CRA from keeping up with its regular audits and lifecycle-based security. This gap can be filled by DYNAMO's dynamic threat intelligence if it is properly integrated. Modern energy systems require real-time monitoring and adaptive response; regulatory frameworks must adapt to support this. Security is a requirement of CRA throughout the product lifecycle, while DYNAMO offers situational awareness and continuity planning. A cybersecurity strategy that is both compliant and operationally sustainable could be created by combining both approaches.

CRA and DYNAMO implementation without integration could result in duplication of effort, inefficiencies, or coverage gaps for organisations. To ensure operational safety, a standardised approach is required to ensure regulatory compliance supports it instead of undermining it. Legal accountability and standardisation across the EU are ensured by the compliance-oriented structure of the CRA. The dynamic and complex threat landscape of the energy sector, particularly in operational technology (OT) environments, may be limited by its static nature and IT-centric focus. DYNAMO's approach is more adaptable and resilient, and it combines business continuity management (BCM) and cyber threat intelligence (CTI) to provide a real-time situational picture. The difference demonstrates the necessity for regulatory frameworks to expand beyond routine audits to include continuous threat monitoring and adaptive risk management.

CRA requires lifecycle security and post-market surveillance to keep product integrity intact. DYNAMO's functionality is strong, but it lacks formal mechanisms to guarantee lifecycle compliance and structured vulnerability disclosure. Energy organisations that exclusively use DYNAMO may face difficulties in meeting CRA's statutory obligations. This could lead to regulatory sanctions or reputational risks due to this difference. The CRA's limited focus on OT systems poses a significant risk because the energy sector's legacy infrastructure frequently lacks modern security features. Although DYNAMO has the potential to address OT-specific threats, further development is necessary to meet CRA's expectations. DYNAMO would benefit from the integration of OT-specific modules and reporting protocols, which would improve regulatory compliance and operational efficiency.

Fragmented implementation of CRA and DYNAMO may lead to inefficiencies or gaps in cybersecurity coverage. A harmonised approach that combines the regulatory rigour of the CRA with the operational agility of DYNAMO would provide a comprehensive cybersecurity strategy that would ensure both compliance and resilience in critical energy infrastructures.

5. Discussion and Conclusions

5.1 Summary of Results

The digitalisation of the energy sector and the shift to more digital systems have made cybersecurity resilience a more important part of comprehensive security. The energy sector forms a high-risk target that is crucial to modern life. Attacks against Ukraine’s electricity grid are an example of the threat that has become possible with digitalisation. CRA is the European Union's legal framework that is developed to help counter these threats.

DYNAMO combines BCM and CTI to enhance organisational resilience. However, DYNAMO lacks some aspects of the CRA framework: DYNAMO should establish a 24-hour reporting protocol to relevant authorities like ENISA, integrate post-market monitoring and develop strategies for OT environments.

A few recommendations should be integrated to align the DYNAMO more closely with the CRA. DYNAMO should integrate structured vulnerability disclosure protocols and 24-hour incident response mechanisms with the European Union Agency for Cybersecurity (ENISA). These core requirements under the CRA ensure rapid identification, communication, and resolution of cyber threats. DYNAMO would enhance its responsiveness and support a coordinated cybersecurity posture across critical infrastructure sectors by embedding these elements into its operational procedures.

There is a need for DYNAMO to address the specific cybersecurity challenges of operational technology (OT) environments and legacy systems, which are particularly prevalent in the energy sector (Stouffer *et al.*, 2023). Unlike newer information technology (IT) systems, these infrastructures often lack built-in security features and are difficult to update. Therefore, DYNAMO should develop tailored strategies that include retrofitting outdated systems, enhancing OT-specific threat detection, and implementing risk controls suited to hybrid IT-OT environments.

In addition, DYNAMO should incorporate active post-market monitoring throughout the lifecycle of digital products. This involves continuous evaluation of security status, timely deployment of updates, and proactive risk management even after products are deployed. Such practices are essential to meeting CRA’s expectations for lifecycle security and to maintain a resilient security posture over time.

Manufacturers must, without undue delay and in any event within 24 hours of becoming aware of it, notify the named authority of any actively exploited vulnerability contained in products with digital elements, as well as any incidents that impact the security of those products (Schmitz-Berndt and Cole, 2022).

5.2 Limitations of the Study

The study's reliance on a qualitative literature review and case study analysis may make its findings less generalizable. The selection and interpretation of publications can be biased when relying on secondary data sources, such as published reports and academic literature. Conceptual gap analysis between the CRA and DYNAMO frameworks is used in the study, but it does not provide quantifiable metrics to evaluate the effectiveness of DYNAMO in real-world compliance scenarios for CRAs. The rapid evolution of cybersecurity threats and regulatory updates can cause some findings to become time-sensitive, making them require ongoing review and adaptation.

5.3 Suggestions for Future Research

Future research should be directed at validating the proposed DYNAMO-CRA compliance strategies through pilot projects in energy sector organisations. Quantitative studies can determine how DYNAMO improvements impact incident response time, vulnerability management, and overall resilience. In addition, cross-sector applications of DYNAMO in healthcare, transportation, and water infrastructure could be explored. To address the dynamic nature of cyber threats, it would be valuable to investigate AI-based threat detection and automated compliance monitoring.

Acknowledgements

This study has received funding from the European Union project DYNAMO, the European Union's Horizon 2020 research and innovation programme under the grant agreement no. 101069601. The views expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Ethics declaration: Ethical clearance was not required for the research.

AI declaration: The paper's spelling was verified using the artificial intelligence tool.

References

DYNAMO project (2024) Dynamic Resilience Assessment Method Including Combined Business Continuity Management and Cyber Threat Intelligence solution for Critical Sectors. Available at: https://horizon-dynamo.eu/wp-content/uploads/2023/01/DYNAMO_Leaflet_web.pdf (Accessed: 24 January 2024).

ECSO (2018) Energy Networks and Smart Grids: Cyber security for the energy sector. Analysis November 2018. Brussels, Belgium: European Cyber Security Organisation, p. 19. Available at: https://ecs-org.eu/ecso-uploads/2022/10/5fdb2673903c6.pdf.

Ene, I.-E. and Savu, D. (2023) 'Cybersecurity - A Permanent Challenge for the Energy Sector', Romanian Cyber Security Journal, 5(1), pp. 107–119. Available at: https://doi.org/10.54851/v5i1y202310.

European Commission. Joint Research Centre (2024) Cyber resilience act requirements standards mapping: Joint Research Centre & ENISA joint analysis. Analysis EUR 31892. Luxembourg: Publications Office of the European Union, p. 69. Available at: https://data.europa.eu/doi/10.2760/905934 (Accessed: 12 August 2025).

European Parliament and of the Council (2024) Cyber Resilience Act. Regulation 2024/2847. European Union, p. 81. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202402847.

ISO (2022) 'ISO/IEC 27001 Standard – Information Security Management Systems'. Geneva, Switzerland: ISO/IEC. Available at: https://www.iso.org/standard/27001 (Accessed: 6 June 2023).

Kaminaris, S. (2025) Cyber Resilience Act: Horizontal cybersecurity requirements for products with digital elements. Available at: https://www.ey.com/en_gr/technical/tax/tax-alerts/cyber-resilience-act-jan-2025 (Accessed: 12 August 2025).

Kim, H., Sefcik, J.S. and Bradway, C. (2017) 'Characteristics of Qualitative Descriptive Studies: A Systematic Review', Research in Nursing & Health, 40(1), pp. 23–42. Available at: https://doi.org/10.1002/nur.21768.

Lee, R.M., Assante, M.J. and Conway, T. (2016) Analysis of the Cyber Attack on the Ukrainian Power Grid. Analysis. Washington, D.C.: E-ISAC, p. 30. Available at: https://nsarchive.gwu.edu/sites/default/files/documents/3891751/SANS-and-Electricity-Information-Sharing-and.pdf.

National Institute of Standards and Technology (2018) Framework for Improving Critical Infrastructure Cybersecurity. Framework. National Institute of Standards and Technology, p. 55. Available at: https://doi.org/10.6028/NIST.CSWP.04162018 (Accessed: 7 June 2023).

Ruoslahti, H. and Tikanmäki, I. (2022) 'E-Skills in Cybersecurity', in A. Dziech, W. Mees, and M. Niemiec (eds) Multimedia Communications, Services and Security (MCSS 2022). Cham: Springer International Publishing, pp. 36–48. Available at: https://doi.org/10.1007/978-3-031-20215-5_4.

Saeed, S. et al. (2023) 'A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience', Sensors (Basel, Switzerland), 23(16), p. 7273. Available at: https://doi.org/10.3390/s23167273.

Schmitz-Berndt, S. and Cole, M. (2022) 'Towards an Efficient and Coherent Regulatory Framework on Cybersecurity in the EU: The Proposals for a NIS 2.0 Directive and a Cyber Resilience Act', Applied Cybersecurity & Internet Governance, 1(1), pp. 1–17. Available at: https://doi.org/10.5604/01.3001.0016.1323.

Shaffique, M.R. (2024) 'Cyber Resilience Act 2022: A silver bullet for cybersecurity of IoT devices or a shot in the dark?', Computer Law & Security Review, 54(September 2024), p. 106009. Available at: https://doi.org/10.1016/j.clsr.2024.106009.

Stouffer, K. et al. (2023) Guide to Operational Technology (OT) Security. Special Publication SP 800-82r3. Gaithersburg, MD: National Institute of Standards and Technology, p. 316. Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-82r3.pdf.

Szedlak, C., Reinemann, H. and Hatzelmann, S. (2024) 'Ensuring Cybersecurity Compliance: Assessing SME Awareness and Preparedness for the Cyber Resilience Act', in Proceedings of Conference on Industrial Engineering and Operations Management. 5th Asia Pacific Conference on Industrial Engineering and Operations Management, Tokyo, Japan: IEOM Society International, USA, p. 10. Available at: https://doi.org/10.46254/AP05.20240160.

World Economic Forum (2025) Fostering Effective Energy Transition 2025. Insight Report June 2025. Cologny/Geneva Switzerland: World Economic Forum, p. 71. Available at: https://reports.weforum.org/docs/WEF_Fostering_Effective_Energy_Transition_2025.pdf.

Yin, R.K. (2009) Case study research: Design and methods. 4th edn. Thousand Oaks, California: SAGE Publications Ltd.

The Life Cycle Approach to Effective Crisis Communications in Mitigating Cyber Threats and Attacks

Jussi Toivanen and Vilma Luoma-aho
University of Jyväskylä, Finland

Jussi.t.toivanen@student.jyu.fi
Vilma.luoma-aho@jyu.fi

Abstract: Cyberattacks pose multifaceted risks, including physical, digital, economic, reputational and societal harm. Often starting as an attack on an organization's intangible assets but quickly spreading to endanger tangible assets, cyberattacks are increasingly common globally. Once an attack becomes public, the targeted organizations face intense scrutiny from employees, customers, the media, the wider public and the authorities. At this point, effective crisis communication becomes critical for mitigating damage and recovering from the crisis. Decision-making and communication during cyber incidents are complicated by uncertainty, operational issues, incomplete information, and even the potential for attackers to shape narratives through extortion or leaks. This paper examines communication challenges in cyber crises through the lens of framing theory and situational crisis communication theory (SCCT). Based on data analysed from interviews with key experts of eight organizations during recent Finnish cyberattacks, the paper identifies key factors that constitute effective cyber crisis communication. These include (1) organizational preparedness, (2) efficient information sharing and coordination, (3) a deep understanding of target audiences and the information environment, and (4) the essential role of communication in leadership. The findings highlight the importance of proactive, transparent and coordinated communication to maintain trust and boost organizational resilience. Practical recommendations in the form of three "no-L-principles" are provided: 1. Do not lessen, 2. Do not lie, and 3. Do not linger. Practical recommendations are provided to enhance crisis communications preparedness, improve real-time crisis communication, reduce risks, and strengthen organizational capacity to manage cyber crises.

Keywords: Cyber crises, Cyberattacks, Resilience, Communication, Crisis communication, Framing, Preparedness

1. Introduction

Vastaamo ("response place" in Finnish) was a Finnish psychotherapy business that became the target of a large-scale data breach in 2020 and filed for bankruptcy in 2021. The hacker managed to obtain the centre's entire database, which contained highly sensitive information on 33,086 clients (Poliisi, 2023). He extorted both the organization and clients by threatening to publish the patient records unless his financial demands were met. The stolen database was published on the dark web, casting a lifelong shadow for victims (Kleeman, 2026). For the organization Vastaamo, the hack was its death strike, with the legal process still ongoing in 2026.

While some cyberattacks may serve military or political objectives (Li and Liu, 2021, p. 8176), often their purpose is disruption or financial gain (Europol, 2025, p. 6, 11). Examples include hacking, ransomware, malware, distribute denial of service (DDoS) attacks, phishing, scams, data breaches and financial fraud (Lallie et al., 2021). The actors behind cyberattacks are multiple, ranging from state actors to organized groups and even individuals (Desa, Juremi and Jenalis, 2025, p. 23). Cyber incidents can trigger many different types of crises, with damage that may remain limited to the digital realm, affecting data, systems and networks but can also extend beyond cyberspace, impacting people, institutions and physical infrastructure (Berg and Kuipers, 2022, p. 21). In severe cases, disruptions can impact international supply chains and escalate into national supply security risks (Turvallisuuskomitea, 2025). Cyber crises can be highly complex and prolonged, potentially accompanied by sustained intensity (Backman, 2020, p. 436). Beyond financial losses, cyber-attacks can erode trust and reputation and harm societies by disrupting social unity, inflicting psychological trauma, and fostering division (Shandler and Gomez, 2023, p.1, 369).

In today's digital environment, the capacity to safeguard consumer and stakeholder data, uphold privacy and maintain strong cybersecurity is increasingly seen as a clear marker of broader social responsibility (Khan, Lee and Liu, 2025, p. 1371). For organizations under attack, attacks are often handled through crisis communication to mitigate the crisis and comply with legal restrictions and requirements (see, e.g., EU NIS2, DORA and GDPR). The organization must ensure that victims receive timely and accurate information without compromising investigative processes or exposing additional vulnerabilities (Traficom, 2025), p. 39). Miller and Pearson (2024) state that in a cyber crisis, the clock starts ticking once the attack has been detected. We argue that from the organization's perspective, the clock actually starts ticking the moment the vulnerability enabling the attack emerges.

In the management of cyber crises, communication plays a central role. The paper asks: What kind of crisis communication is needed to manage cyberattacks? Building on Knight and Nurse (Knight and Nurse, 2020) and the process-like nature of cyber incidents, we structure this paper into three stages: before, during and after crisis communication.

2. Organizational Cyber Crisis Communications

As everyday life becomes more deeply intertwined with cyberspace, even small cyber incidents can lead to significant and wide-ranging crises (Berg and Kuipers, 2022, p. 21). The uncertainty created by a crisis places psychological strain on stakeholders. This strain can be reduced by providing timely, clear information about what happened and what steps will be taken to prevent a similar crisis in the future. (Coombs, 2007, p. 165). Crisis communication has been defined as "the ongoing process of creating shared meaning of among and between groups, communities, individuals and agencies within the ecological context of a crisis for the purpose of preparing for and reducing, limiting, and responding to threats and harm" (Sellnow and Seeger, 2021, p. 17).

When an organization becomes a target, it immediately forfeits part of the symbolic capital that reputation represents (Perera et al. 2022, p. 2). In public discourse, attacks are judged not only by technical details but by the narrative built around them. Communication becomes a crucial part of the impact chain: it can soften the crisis or intensify it. Cyber incidents are demanding because information is often incomplete, fragmented and unreliable (Traficom, 2025). The situation becomes clear only as the investigation progresses, so communication is built largely on uncertain information (Traficom, 2025). Silence can be harmful: if an organization fails to communicate clearly, stakeholders and the media may perceive secrecy or guilt, damaging organizational reputation (Pang, 2013, cited in Woon and Pang, 2017, p. 331). Therefore, it is crucial for the organization to communicate promptly about what has occurred, even if all the details are not yet known, often balancing uncertainty with swift communication (Knight and Nurse, 2020, p. 13).

Previous studies on cyber crises and communications have explored various topics – including the impact of data breach announcements on stock prices (Rosati et al., 2017), the role of corporate reputation and crisis response strategies in managing data breaches (Gwebu, Wang and Wang, 2018), resilience and its significance in managing cyber incidents (Calder, 2023; Joinson et al., 2023; Tzavara and Vassiliadis, 2024), testing perceptions on organizational apologies after a data breach (Bentley and Ma, 2020) – and provided behavioural recommendations (Zhang and Borden, 2020). Further, studies (Agrafiotis et al., 2018) have categorized cyber-related long-term harms around five key themes: physical or digital harm, economic harm, psychological harm, reputational harm and social harm. They define cyber harm as "the damage that arises as a direct result of an attack conducted wholly or partially via digital infrastructures, and the information, devices and software applications that these infrastructures are composed of" (Agrafiotis et al., 2018, p. 2).

From the practice side, Knight and Nurse (2020) propose a cyber-crisis communication framework for the pre-crisis and during-crisis phases, drawing on an extensive literature review and expert interviews that provide an empirical foundation for their recommendations on preparedness and communication. In addition to research-based guidance, practical CERT/CSIRT-oriented guidelines also support preparedness, response and recovery, for example, NIST's widely used Incident Response Recommendations (SP 800-61), whose third updated version was published in April 2025 (Nelson et al., 2025). However, these technically oriented guidelines tend to approach communication narrowly and often only superficially. When communication is viewed merely as support for technical functions and operational recovery, crisis management suffers because the organization overlooks the human dimension at the core of crises – how people interpret events and how they respond.

3. Theoretical Framework

Cybersecurity is not just about technology. It also involves the people who use technology and ensure it is implemented and operated correctly (Chaudhary, Gkioulos and Katsikas, 2022, p. 1). Research on cybersecurity spans multiple disciplines, ranging from technology and law to psychology and sociology (Fujs, Mihelič and Vrhovec, 2019, p. 1). An important question is where the focus of cybersecurity research should be directed. A human and social science perspective (Fujs, Mihelič and Vrhovec, 2019) on this technological topic is valuable, as it enables a broad analysis – both quantitative and qualitative – of the multidimensional human-technical context and offers insights accordingly.

As Triplett (2022, p. 575) states, "People and cybersecurity are inseparable". Therefore, communication constitutes a core capability for building and maintaining organizational cybersecurity (Triplett, 2022, p. 575) and for conveying attacks, their impacts and action guidance to employees and potential victims (Knight and

Nurse, 2020). Crises and crisis management likewise concern organizational reputation, conditioned by how employees, stakeholders and the media interpret and make sense of events (Coombs, 2007, p. 164). Accordingly, organizations should recognize that crisis situations and their management are not only operational processes but also intrinsically communicative. Also, as Coombs, Holladay and White (2020, p. 41) point out, effective crisis responses need to demonstrate empathy towards those affected, as showing empathy helps build trust between crisis managers and their stakeholders.

The growing use of digital channels, especially social media, has created new challenges for effective crisis communication (Coombs, Holladay and White, 2020, p. 35). Also, numerous actors seek to shape the information environment for their own purposes, and digital technologies give them capabilities that far surpass the public's understanding of what is possible or already underway (Wanless and Pamment, 2019, p. 1). At the same time, classical computer-mediated communication is expanding into artificial intelligence–mediated communication (AI-MC), in which AI is not merely a conduit but an active participant that can modify, augment or even generate messages so that people can better achieve their communicative goals (Hancock, Naaman and Levy, 2020, p. 90–91). Against this backdrop, it becomes crucial in crises to understand how public frames are constructed and processed (van der Meer, 2018) and how an organization's own actions and communication influence stakeholder interpretations (Coombs, 2007, p. 171).

Therefore, organizations must recognize that cyberattacks, despite their technical nature, function as communicative acts that shape the information environment. Regardless of the attacker's motives, their actions inevitably convey a message (Traficom, 2025). This means that cyber incidents must be interpreted not only as technical disruptions but also as meaning-making events that influence perceptions, trust and stakeholder expectations.

3.1 Framing the Cyber Incident

One challenge in crisis communication is that people interpret events differently, so messages may be misunderstood or tied to unrelated narratives despite the organization's intent, known as framing. Chong and Druckman (2007, p. 104) define framing as "the process by which people develop a particular conceptualization of an issue or reorient their thinking about an issue".

In contemporary digital societies, individuals can access, generate, circulate and reinterpret information through a wide range of online platforms, reshaping how meaning is constructed (Johansson et al., 2024; Neuman et al., 2014). As a result, framing becomes more dynamic and increasingly difficult for organizations to manage. This difficulty is further intensified by the digitalization of networked framing processes, which significantly influence interpersonal and organizational communication across both online and offline environments. These dynamics shape how groups emerge and divide, process new information, and remain vulnerable to misinformation. (Entman and Usher, 2018, p. 299). From the perspective of planning and implementing crisis communication it is essential for organizations to recognize and understand the diversity of framing and the ways in which frames emerge and evolve within a digital society.

According to media framing theory, the media can shape how the public interprets complex issues by selectively emphasizing certain elements (Zhang and Zhang, 2024, p. 13). In the era of social media, traditional media still retains an important role. As Roulet and Clemente (2018, p. 328) describe the role of traditional media as producing common knowledge about various events. In the context of social media, people then interact and respond on the basis of this shared knowledge. This process creates a continuous interplay between traditional media and social media.

Snider et al. (2021, p. 2) point out that past cases have shown that in many cyberattacks, the public internalizes the details of the attack immediately after it occurs, when media coverage is at its peak. Even though more accurate information is released later, fear and concern have already taken hold. From an organizational perspective, it is essential to provide sufficient contextual information when communicating about an event. Furthermore, according to White (2009, p. 187), when the media report issues without organizational context, negative framing tends to intensify, elevating issue salience and perceptions of crisis severity. If issues are not managed proactively, external audiences often adopt the media's perspective. Finally, in crises, the way an organization shapes its communication and provides contextual information can influence how the media frames and portrays a crisis (White, 2009, p. 178).

3.2 Situational Theory of (Cyber) Crisis Communication

The importance of communication becomes especially clear in times of crisis, when organizations must respond quickly, maintain trust and manage uncertainty. Coombs and Holladay (2022, p. 259–260) emphasize that crisis communication is a strategic tool that enables organizations to adapt to disruption, protect stakeholders and regain focus on core objectives. Whereas crisis management refers to the actions taken to address a crisis, crisis communication is the mechanism through which those actions are carried out and made meaningful to those affected (Coombs and Holladay, 2022, p. 259-260).

Widely used in organizational crisis communication research, situational crisis communication theory (SCCT) provides a model for anticipating how stakeholders will react and what kind of reputational threat the crisis may pose (Coombs, 2007). The theory highlights how key aspects of a crisis situation influence attributions related to the crisis and stakeholders' perceptions of the organization's reputation (Coombs, 2007). This is important in managing crises.

From this perspective, according to Coombs (2007, p. 166), three factors influence the reputational impact of a crisis: how responsible the organization is perceived to be, whether it has experienced previous crises, and what its reputation was like before the crisis. The more responsibility and negative history the organization has, the greater the reputational threat. SCCT highlights the importance of perceived responsibility and how it shapes reputation. Knight and Nurse (2020, p. 15) note that in the context of data breaches, organizations may be inclined based on SCCT to portray attackers or other external actors as those at fault, thereby positioning themselves primarily as victims.

Drawing from recent literature and real-world cases in Finland, this paper applies framing theory and situational crisis communication theory to analyse organizational crisis communication and its challenges concerning cyberattacks.

Research questions

The paper states that clear crisis communication strategies can be identified that organizations have used to manage and mitigate cyber incidents. These strategies are shaped by the specific characteristics and severity of the case, the organization's own features, its level of preparedness, general crisis readiness, previous crisis history, and the leadership's understanding of cyber threats. This paper analyses these approaches and is guided by the research question: "How can effective crisis communication be used to mitigate cyber risks and threats?" These insights aim to enhance organizational preparedness and crisis leadership from a crisis communications perspective.

4. Research Method

Multiple challenges relate to cyber security studies. In cybersecurity contexts, the target of an attack may be reluctant to disclose details about the incident due to concerns about operational issues, reputation, legal implications or the challenges of attribution (Lilli, 2023, p. 6). The data collection for this paper also involved challenges, which are described below.

For this study, a qualitative, semi-structured thematic interview method was selected to explore events and their underlying factors in depth, as the method enables participants from a limited number of organizations to share their personal perspectives, interpretations and meaning (Tuomi and Sarajärvi, 2018).

Grounded in SCCT and framing theory, the developed interview guide covered the full crisis life cycle through thematic sections, assessing preparedness, leadership and situational management. Questions related to communication practices, constraints, reputation, trust, resilience and accountability were asked. It also examined public framing and included questions on post-crisis recovery, communication and evaluation. Follow-up questions were asked to deepen understanding based on the interviewees' responses.

The study examined Finnish public- and private-sector organizations hit by high-visibility cyberattacks since 2020. Incidents spanned DDoS, phishing/scam campaigns, data breaches and ransomware, with several organizations facing recurring attacks, often repeated phishing and DDos, over subsequent years. The interviews focused on the attack types most relevant to each organization.

Experts were selected based on their roles and incident-management involvement, identified either directly or through organizational referrals. Personnel changes sometimes required redirecting interviews to those with first-hand experience. Interviewees typically prepared by consulting internally with CISOs and other specialists involved in the response.

Despite varied job titles, all interviewees participated in planning or executing crisis communication for cyber incidents. Crisis-team structures differed across organizations, reflecting organizational design and role allocations between communication and cybersecurity experts.

Table 1: Interviewed organizations, sector and expert roles, and attack types

Organization	1	2	3	4	5	6	7	8
Attack type	Ransom ware	Data breach and ransomwar e	Data breach and ransomwar e	Data breach	Data breach and ransom ware	DDoS /scams/Phi shing	Ddos/scams/P hishing	Data breach and ransomwar e
Intervie wed	Technical Director, Administr ative Director and occupatio nal safety officer	Director of communic ations	CIO and vice president (global marketing and communica tions)	Communic ations director	Chief	Communic ations director and communic ations manager	Corporate and customer communicatio ns lead	Head of security and external communic ations manager
Sector	Public	Public	Private	Public	Private	Private	Private	Private

Seventeen organizations were approached for interview participation. Two declined at the initial contact. One organization initially consented but did not engage with subsequent scheduling communications. Three organizations did not respond to either the initial invitation or the follow-up message. One organization initially accepted but cancelled and deferred the interview upon receipt of the advance questions and did not subsequently reinitiate scheduling. One organization redirected the request to an external communications agency that had supported the organization's crisis communication during the cyberattack. For one organization, an interview could not be arranged within the data-collection period due to scheduling constraints.

Participants received the interview questions in advance and retained the right to decline specific items or withdraw entirely. The interviews lasted 45–120 minutes and were conducted primarily via Zoom or Teams, with one held in person. All sessions were recorded and stored securely in encrypted form on the University of Jyväskylä's network drive and then deleted after transcription. During transcription, anonymization procedures ensured that neither organizations nor individual experts could be identified.

Given the interviewer's cybersecurity background and prior familiarity with some cases and interviewees, special care was taken to ensure that participants could speak freely and steer the discussion without undue influence. To preserve anonymity, the industries of the organizations are not reported, as such details could enable identification, given the small number of publicly known cyberattacks in Finland. Safeguarding anonymity was a central consideration throughout the study.

Following the interviews, the material was transcribed and analysed using the inductive method (Thomas, 2006). An inductive approach was selected, as it was considered particularly well suited for analysing interview data. Instead of relying on predefined theories, the aim was to allow the material itself to guide the process, identifying naturally emerging themes and categories directly from the transcripts. These insights could then be used to construct a framework that reflects the participants' perspectives and experiences (Thomas, 2006, p. 238).

5. Cyber Crisis Communications

5.1 Communication Before a Cyber Crisis

Proactive communication and building awareness are essential when preparing for and responding to cyberattacks. Based on the interviews, proactive communication helps employees and external parties understand how the organization prepares for and responds to cyberattacks. It also provides customers and partners with guidance for safeguarding their own data and devices. Such communication strengthens trust and improves the ability to withstand and recover from cyber crisis (Hytönen and Ruoslahti, 2025). Investing in intangible assets before any crisis is a key means of building resilience and securing support when disruption occurs (Canel, Luoma-aho and Barandiarán, 2020, p. 171–172).

Cyber incidents often remain undetected for extended periods, allowing vulnerabilities to persist until an attack or accident escalates into a full-scale crisis, leaving decision makers little time to prepare or respond (Berg and Kuipers, 2022, p. 14). High uncertainty and ambiguity further complicate management: attacks can originate anywhere, unfold instantly and produce immediate effects, unlike more predictable threats with warning signs. The lack of early indicators hinders prevention and complicates both preparedness and effective crisis response (Berg and Kuipers, 2022, p. 14).

The interview data indicate that effective pre-incident communication entails the following: (i) mapping plausible cyber-threat scenarios and associated impacts; (ii) identifying priority audiences; (iii) integrating and analysing crisis history and existing trust capital; and (iv) planning across the entire threat life cycle, including post-incident recovery. To mitigate uncertainty, (i) organizations should draft scenario-specific frameworks that specify management and information (situational-picture) requirements at each organizational tier, (ii) bring external partners into case-specific preparedness planning, (iii) and pre-draft communications for different channels and audiences across likely scenarios. Given the likelihood of a prolonged disruption, respondents emphasized the following: (i) redundant channels, (ii) clearly assigned management and communications roles with deputies, (iii) offline storage of essential materials, and (iv) processes that are deliberately designed, documented and regularly exercised.

Awareness-building is likewise central to preparedness. Effective communication improves risk comprehension (Chaudhary, Gkioulos and Katsikas, 2022), as underscored in Interviews 6 and 7. By proactively educating staff and customers, organizations can build trust and strengthen their readiness. Routine preparedness messaging also shapes stakeholder perceptions during a cyberattack (Interviews 6 and 7).

5.2 Communication During a Cyber Crisis

Interviews suggest that the early phase of cyber incidents is often marked by acute uncertainty: organisations may lack a consolidated situational picture - what happened, the scope, and which internal or external communication channels are safe to use. Technical constraints, legal obligations and ongoing investigations can restrict disclosure, creating significant challenges for determining what can be communicated, to whom, and at what point in time.

Effective crisis communication was generally seen as contingent on robust preparedness, clearly communicated command and management structures, regular exercises and dedicated expertise; it also benefits from alignment with stakeholder information needs and awareness of the broader information environment. Timely, level-specific and tailored situational picture was seen important for leadership, operational activities and communication.

Interviewees emphasised early stakeholder identification and the value of explicit communication coordination mechanisms; roles, responsibilities, and deputies are typically best defined and communicated in advance so that tasks, authority, and decision rights are clearer. It was considered important that internal personnel be informed first to support consistent responses, effective communication and customer care. In addition, worst-case contingency planning, including the possibility of a full data leak, was advised early on to keep messaging timely, coherent, and relevant amid evolving information. In parallel, organisations were advised to start planning and preparation for recovery communication, even while the investigation and corrective actions were in progress.

From the perspective of organizations, crises invariably raise questions of responsibility and reputation (Coombs, 2007) and trust (Siegrist, Gutscher and Keller, 2010). Beyond direct financial losses, cyberattacks cause significant intangible damage to morale and goodwill, which are far more difficult to measure (Perera et al., 2022, p. 2). Restoring trust requires transparent, evidence-based communication that does not downplay the incident (Knight and Nurse, 2020). Based on the interviews, common communication errors included evading responsibility, withholding information, minimising events, or issuing false statements. Existing trust capital, reputation and an organization’s crisis history were found to influence how audiences interpret events. It was also recommended that communication planning should consider concurrent events and stakeholder expectations. Organizations should also stay aware of the broader context and ongoing developments in the information environment to ensure that the timing, content, and tone of their communications are properly aligned. This helps preserve credibility, reduce ambiguity, and support reputational recovery.

5.3 Communication After a Cyber Crisis

Interviews showed that organizations' view post-incident communication as essential for employees, possible victims, customers and partners. Openness was the preferred strategy to rebuild trust and reputation, and one expert interviewed felt that sharing lessons learned during and after an attack generated positive feedback and helped restore trust and credibility (Interview 3).

Cyber crises can have lasting effects on personnel and operations. Reviewing the incident, especially with those most involved, was deemed important for learning and recovery. Staff may remain sensitive to minor technical issues, as shown by an organization recovering from ransomware (Interview 1). Operational impacts can persist, for example, when lost data cannot be restored (Interview 1), making it necessary to issue targeted communications that provide clear guidance and support to employees. Acknowledging employees' efforts after the incident was also considered important (Interview 2).

As one expert noted (Interview 5) crisis communication reveals more about an organization than any branding in normal circumstances because values, strategy and identity are tested in practice. Effective crisis communication therefore depends on shared values and a clear strategic direction, which enable responsible and coherent communication under pressure.

6. Recommendations for Managing Cyber Crisis Communications

Based on interviews and research literature, organizations are advised to use a life cycle approach to cyber crisis communication. Plan, implement and review actions across the preparedness, response and recovery phases. Clear roles, responsibilities and procedures with effective information flow are associated with more effective communication and leadership during incidents.

After a crisis, communication should continue with timely updates and audience-specific follow-ups for staff, affected individuals, customers and partners, acknowledging their different information needs. The next section explains how communication can help mitigate cyber threats and manage attacks.

The following recommendations are organized into (i) leadership, (ii) communication, and (iii) situational awareness. They align with the Finnish government's crisis communication guidelines (Prime Minister's Office, 2019), which were developed from post-analyses of crises affecting Finnish society since 2012 and emphasize the seamless integration of leadership, communication with effective information flow. This interconnection was further reinforced by the Finnish government's 2021 crisis communication development project (Valtioneuvoston kanslia, 2023), informed by broad stakeholder consultations and data collection.

Before the crisis

Leadership: Organizations should ensure preparedness and continuity planning for cyber threats, including activating a diverse crisis management team with clear roles, responsibilities and backup arrangements. Partners and relevant stakeholders must also be identified, cooperation should be agreed upon in advance, and coordination structures must be established.

Communications: Crisis communication plans must cover different attack types, identify spokespersons and substitutes, define core messages and prepare channel-specific communication materials in advance for various audiences. Backup communication channels should be tested and ready. Partners and relevant stakeholders must also be identified, cooperation should be agreed upon in advance, and coordination structures must be established.

Situational awareness: Processes for gathering and verifying information should be established, including alternative methods for maintaining situational awareness when systems fail. Plans should define what information is essential for decision-making and operations in different scenarios and for each level of the organization.

During the crisis

Leadership: The crisis team should be activated immediately to manage operations, ensure compliance, and maintain continuity. Leadership must coordinate decision-making and support recovery actions. During a crisis, management must consider the personnel's operational capacity, including their psychological well-being, by demonstrating support and understanding and by maintaining a visible presence. Leadership should take responsibility and remain visibly engaged with external stakeholders.

Communications: Communication should be timely, accurate and consistent across channels. Employees should be informed first, and designated spokespersons should deliver messages supported with prepared materials. Backup channels should be used if primary systems are compromised. Communication coordination networks must be activated, and the flow of information between them must be ensured. The communications function must have access to an up-to-date operational situational picture.

Situational awareness: Organizations must identify the attack type, scope and impact quickly, update the situational picture continuously, and share it with relevant audiences. Monitoring internal and external impacts and supporting leadership decisions are critical. The flow of situational awareness information must be ensured between all actors and organizations involved in the situation.

After the crisis

Leadership: Assess the impact on operations and stakeholders, implement recovery measures, and document lessons learned for future preparedness. Attention should be given to restoring normal operations and supporting staff in the recovery process. Senior management must engage in active stakeholder communication.

Communication: Post-crisis communication should demonstrate accountability and inform stakeholders about actions taken. As the situation is clarified and the investigation progresses, the instructions are refined as needed. Sharing communication lessons learned internally and externally supports resilience.

Situational awareness: Continue monitoring direct and indirect impacts. Analyse the incident to compile lessons learned and improve future situational awareness processes.

7. Conclusion

Cyberattacks disrupt organisations' operations, decision-making, and communication. Interviews across sectors and attack types revealed consistent themes: managing uncertainty, addressing fear, communicating with incomplete information, safeguarding continuity, and protecting intangible assets such as reputation and trust. As highlighted in the interviews, victims' and customers' perspectives must be considered, and guidance should be clear, practical, and action-oriented. Irrespective of the attack type, customers and other external stakeholders consistently expressed concern that sensitive data might fall into the hands of attackers.

According to the interviews, it was notable that, regardless of attack type, the same kinds of concerns consistently emerged. This suggests that variations in attack modalities did not significantly alter the core requirements for crisis communication. Across contexts and cases, the fundamental communication practices remained largely the same. We took this observation into account when developing our framework: the general principles we identified are applicable across different types of cyberattacks.

Our findings extend the framework proposed by Knight and Nurse (2020) by specifying a distinct post-crisis phase in which communication continues to play a central role: providing updates, supporting those affected, rebuilding trust, and strengthening resilience. Investing in post-crisis communication is essential, as it plays a vital role in ensuring organisational continuity, recovery, and the long-term restoration of reputation and trust. Employees also emerged as a distinct audience requiring focused attention in preparedness and in every crisis phase.

Interviews highlighted that cyber crises place heavy demands on managers, requiring situational leadership in fast-changing conditions. Communication was seen as a crucial enabler, with supervisors playing a key role as the closest link to employees and the frontline. Overall, the interviews underscored that communication is a core leadership task in cyber crises: open, empathetic, two-way dialogue sustains trust, reduces uncertainty, supports wellbeing, and strengthens organisational resilience.

Based on the interviews, cyber crisis communication can be summarized into the three "no-L principles": do not lessen, do not lie, and do not linger. The idea is that in a crisis situation, an organization must be honest, direct, and timely, without trying to downplay or hide the problem. This builds trust and enhances the organization's ability to manage and recover from the crisis more effectively.

AI declaration: Generative AI was utilized to improve the clarity and language of the text. Following this, the author thoroughly reviewed and revised the content to ensure its accuracy, taking full responsibility for the final version.

Ethics declaration: The ethical guidelines of the University of Jyväskylä were followed throughout the study.

References

Agrafiotis, I. et al. (2018) 'A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate', *Journal of Cybersecurity*, 4(1), p. tyy006. Available at: https://doi.org/10.1093/cybsec/tyy006.

Backman, S. (2020) 'Conceptualizing cyber crises', *Journal of Contingencies and Crisis Management*, 29. Available at: https://doi.org/10.1111/1468-5973.12347.

Bentley, J.M. and Ma, L. (2020) 'Testing perceptions of organizational apologies after a data breach crisis', *Public Relations Review*, 46(5), p. 101975. Available at: https://doi.org/10.1016/j.pubrev.2020.101975.

Berg, B. van den and Kuipers, S. (2022) 'Vulnerabilities and cyberspace: A new kind of crises', in *Oxford Research Encyclopedia of Politics*. Available at: https://doi.org/10.1093/acrefore/9780190228637.013.1604.

Calder, A. (2023) *Cyber resilience: Defence-in-depth principles*. First edition. Ely: ITGP. Available at: https://research.ebsco.com/linkprocessor/plink?id=fbb74d3d-f862-37b2-8a4c-e319d5b5bbbd (Accessed: 9 September 2025).

Canel, M.-J., Luoma-aho, V. and Barandiarán, X. (2020) 'Public sector communication and publicly valuable intangible assets', in *The Handbook of Public Sector Communication* (eds V. Luoma-aho and M.-J. Canel). https://doi.org/10.1002/9781119263203.ch6

Chaudhary, S., Gkioulos, V. and Katsikas, S. (2022) 'Developing metrics to assess the effectiveness of cybersecurity awareness program', *Journal of Cybersecurity*, 8(1), tyac006. Available at: https://academic.oup.com/cybersecurity/article/8/1/tyac006/6590603

Chong, D. and Druckman, J. N. (2007) 'Framing theory', *Annual Review of Political Science*, 10, pp. 103–126. Available at: https://fbaum.unc.edu/teaching/articles/Chong-Druckman-FramingTheory.pdf

Coombs, W.T. (2007) 'Protecting Organization Reputations During a Crisis: The Development and Application of Situational Crisis Communication Theory', Corporate Reputation Review, 10(3), pp. 163–176. Available at: https://doi.org/10.1057/palgrave.crr.1550049.

Coombs, W.T., Holladay, S.J. and White, C.L. (2020) The Handbook of Crisis Communication. 2nd edn. Wiley-Blackwell.

Coombs, W.T. and Holladay, S.J., 2022. 'Crisis communication as strategic communication: Process and insights', in J. Falkheimer and M. Heide (eds.) *Research handbook on strategic communication*. Cheltenham: Edward Elgar, pp. 259–273. doi:10.4337/9781800379893.00025

Desa, J.M., Juremi, J. and Jenalis, M.H. (2025) 'Understanding cyber threat actors: A review on classification, motivations, and behavioral profiling', in V. Bhateja et al. (eds.) *Innovations in communication networks: sustainability for societal and industrial impact*. Singapore: Springer Nature, pp. 23–30. Available at: https://doi.org/10.1007/978-981-96-5223-5_3.

Entman, R.M. and Usher, N. (2018) 'Framing in a fractured democracy: Impacts of digital technology on ideology, power and cascading network activation', *Journal of Communication*, 68(2), pp. 298–308. Available at: doi: 10.1093/joc/jqx019.

Europol (2025) *Steal, deal and repeat – How cybercriminals trade and exploit your data: Internet Organised Crime Threat Assessment (IOCTA) 2025*. Luxembourg: Publications Office of the European Union. Available at: https://www.europol.europa.eu/cms/sites/default/files/documents/Steal-deal-repeat-IOCTA_2025.pdf

Falkheimer, J. and Heide, M. (2010) 'Crisis communicators in Change: From plans to improvisations', in *The handbook of crisis communication*. John Wiley & Sons, Ltd, pp. 511–526. Available at: https://doi.org/10.1002/9781444314885.ch25.

Fujs, D., Mihelič, A. and Vrhovec, S.L.R. (2019) 'The power of interpretation: Qualitative methods in cybersecurity research', in Proceedings of the 14th International Conference on Availability, Reliability and Security. ARES '19: 14th International Conference on Availability, Reliability and Security, Canterbury CA United Kingdom: ACM, pp. 1–10. Available at: https://doi.org/10.1145/3339252.3341479.

Gwebu, K.L., Wang, J. and Wang, L. (2018) 'The role of corporate reputation and crisis response strategies in data breach management', *Journal of Management Information Systems*, 35(2), pp. 683–714. Available at: https://doi.org/10.1080/07421222.2018.1451962.

Hancock, J.T., Naaman, M. and Levy, K. (2020). 'AI-mediated communication: Definition, research agenda, and ethical considerations', *Journal of Computer-Mediated Communication*, 25(1), 89–100.

Hytönen, E. and Ruoslahti, H. (2025) 'Crisis communication guidelines to support cyber resilience', *European Conference on Cyber Warfare and Security*, 24, pp. 222–229. Available at: https://doi.org/10.34190/eccws.24.1.3730.

Johansson, S., Johansson, B. and Johansson, J. (2024) 'The dynamics of information-seeking repertoires: A cross-sectional latent class analysis of information-seeking during the COVID-19 pandemic', *Mass Communication and Society*, 27(4), pp. 599–626. Available at: https://doi.org/10.1080/15205436.2023.2258863.

Joinson, A.N. et al. (2023) 'Development of a new "human cyber-resilience scale"', *Journal of Cybersecurity*, 9(1), p. tyad007. Available at: https://doi.org/10.1093/cybsec/tyad007.

Khan, W.N., Lee, J.K. and Liu, S. (2025) 'Is cybersecurity a social responsibility?', *Information Systems Frontiers*, 27(4), pp. 1367–1391. Available at: https://doi.org/10.1007/s10796-024-10565-z.

Kleeman, J. (2026) *A faceless hacker stole my therapy notes – now my deepest secrets are online forever*. BBC. Available at: https://www.bbc.com/news/articles/c62nzxqw45eo

Knight, R. and Nurse, J. (2020) 'A framework for effective corporate communication after cyber security incidents', *Computers & Security*, 99. Available at: https://doi.org/10.1016/j.cose.2020.102036.

Lallie, H.S. et al. (2021) 'Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic', *Computers & Security*, 105, p. 102248. Available at: https://doi.org/10.1016/j.cose.2021.102248.

Li, Y. and Liu, Q. (2021) 'A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments', *Energy Reports*, 7. Available at: https://doi.org/10.1016/j.egyr.2021.08.126.

Lilli, E. (2023) 'How can we know what we think we know about cyber operations?', *Journal of Global Security Studies*, 8(2), p. ogad011. Available at: https://doi.org/10.1093/jogss/ogad011.

Meer, van der T.G.L.A. (2018) 'Public frame building: The role of source usage in times of crisis', *Communication Research*, 45(6), pp. 956–981. Available at: https://doi.org/10.1177/0093650216644027.

Miller, K. and Pearlson, K. (2024) 'How to build a cyber crisis communications plan', *MIT Sloan Management Review*, 16 September. Available at: https://sloanreview.mit.edu/article/how-to-build-a-cyber-crisis-communications-plan/

Nelson, A. et al. (2025) 'Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile', *NIST Special Publication* (SP) 800-61 Rev. 3. National Institute of Standards and Technology. Available at: https://doi.org/10.6028/NIST.SP.800-61r3.

Perera, S., Jin, X. and Opoku, D-G. J. (2022) 'Factors affecting reputational damage to organisations due to cyberattacks', *Informatics*, 9(1), 28. Available at: https://www.mdpi.com/2227-9709/9/1/28

Poliisi (2020) Tietomurron tutkinta jatkuu – poliisi toivoo harkintaa asiaan liittyvien yksityiskohtien julkaisemisessa. Available at: https://poliisi.fi/-/tietomurron-tutkinta-jatkuu-poliisi-toivoo-harkintaa-asiaan-liittyvien-yksityiskohtien-julkaisemisessa-1

Prime Minister's Office (2020) *Guidelines for enhanced government communications: Communications under normal conditions and during incidents*. Publications of the Prime Minister's Office 2020:9. Available at: https://julkaisut.valtioneuvosto.fi/items/735b999e-5579-466b-819d-435794dde21c

Rosati, P. et al. (2017) 'The effect of data breach announcements beyond the stock price: Empirical evidence on market activity', *International Review of Financial Analysis*, 49, pp. 146–154. Available at: https://doi.org/10.1016/j.irfa.2017.01.001

Roulet, T.J. and Clemente, M. (2018) Let's open the media's black box: The media as a set of heterogeneous actors and not only as a homogenous ensemble. *Academy of Management Review*. doi:10.5465/amr.2016.0537

Russell Neuman, W. et al. (2014) 'The dynamics of public attention: Agenda-setting theory meets big data: Dynamics of public attention', *Journal of Communication*, 64(2), pp. 193–214. Available at: https://doi.org/10.1111/jcom.12088

Sellnow, T.L. and Seeger, M.W. (2021) Theorizing crisis communication. John Wiley & Sons.

Shandler, R. and Gomez, M.A. (2023) 'The hidden threat of cyber-attacks – undermining public confidence in government', *Journal of Information Technology & Politics*, 20(4), pp. 359–374. Available at: https://doi.org/10.1080/19331681.2022.2112796

Siegrist, M., Gutscher, H. & Keller, C. (2010) 'Trust and Confidence in Crisis Communication: Three Case Studies', in Earle, T.C. & Gutscher, H. (eds.) Trust and Confidence in Risk Communication. United Kingdom: Taylor & Francis Group. Available at:

Snider, K. L. G., Shandler, R., Zandani, S. and Canetti, D. (2021) Cyberattacks, cyber threats, and attitudes toward cybersecurity policies. *Journal of Cybersecurity*, 7(1), tyab019. Available at: https://academic.oup.com/cybersecurity/article/7/1/tyab019/6382745

Thomas, D.R. (2006) 'A general inductive approach for analyzing qualitative evaluation data', *American Journal of Evaluation*, 27(2), pp. 237–246. Available at: https://doi.org/10.1177/1098214005283748.

Traficom (2025) Miten viestiä kyberhyökkäyksistä? – kriisiviestintäohje organisaatioille. Traficom. Available at: https://traficom.fi/fi/julkaisut/miten-viestia-kyberhyokkayksista-kriisiviestintaohje-organisaatioille

Traficom and Supo (2025) *Cyber security threat level remains high – serious cases on the rise*. Available at: https://www.kyberturvallisuuskeskus.fi/en/news/traficom-and-supo-cyber-security-threat-level-remains-high-serious-cases-rise

Triplett, W.J. (2022) 'Addressing human factors in cybersecurity leadership', *Journal of Cybersecurity and Privacy*, 2(3), pp. 573–586. Available at: https://doi.org/10.3390/jcp2030029.

Turvallisuuskomitea (2025) Yhteiskunnan turvallisuusstrategia. Available at: https://turvallisuuskomitea.fi/yhteiskunnan-turvallisuusstrategia/

Tuomi, J. and Sarajärvi, A. (2018) Laadullinen tutkimus ja sisällönanalyysi. Uudistettu laitos. Helsinki: Kustannusosakeyhtiö Tammi.

Tzavara, V. and Vassiliadis, S. (2024) 'Tracing the evolution of cyber resilience: A historical and conceptual review', *International Journal of Information Security*, 23, pp. 1–25. Available at: https://doi.org/10.1007/s10207-023-00811-x.

Valtioneuvoston kanslia (2023) Kriisitilanteita koskevan kansalaisviestinnän kehittäminen: Työryhmän loppuraportti. Valtioneuvoston kanslian julkaisuja 2023:1. Available at: https://julkaisut.valtioneuvosto.fi/items/7bfd752c-ea0f-43ed-b94f-574e08733ef9

Wanless, A. and Pamment, J. (2019) 'How do you define a problem like influence?', *Journal of Information Warfare*, 18(3), pp. 1–14. Available at: https://www.jinfowar.com/journal-issue/volume-18-issue-3

White, C. (2009) 'Examining a crisis communication void: The role of context to mitigate issues', *Journal of Communication Management*, 13(2), pp. 176–190. Available at: https://doi.org/10.1108/13632540910951777.

Woon, E. and Pang, A. (2017) 'Explicating the information vacuum: Stages, intensifications, and implications', *Corporate Communications: An International Journal*, 22(3), pp. 329–353. Available at: https://doi.org/10.1108/CCIJ-10-2016-0066.

Zhang, Q. and Zhang, X. (2024) 'Media influence on public trust during crises: A comparative analysis of different media types and trust dimensions', *Journal of Contingencies and Crisis Management*, 32(3), p. e12624. Available at: https://doi.org/10.1111/1468-5973.12624.

Zhang, X.A. and Borden, J. (2020) 'How to communicate cyber-risk? An examination of behavioral recommendations in cybersecurity crises', *Journal of Risk Research*, 23(10), pp. 1336–1352. Available at: https://doi.org/10.1080/13669877.2019.1646315

Hybrid Learning Techniques for Image Forensics and Privacy Protection in the Face of Deep Fake Threats

Arif Ullah[1], Sidra Pervez[2], Muhammad Wajidullah Khan[3], Aina Hassan[4]and Muazam Ali[2]
[1]Faculty of Computing and Artificial Intelligence, Air University, Islamabad, Pakistan
[2]Department of Management Sciences, HITEC University, Taxila, Pakistan
[3]Faculty of Engineering and Technology UTHM, Malaysia
[4]Othman Yeop Abdullah Graduate School of Business, UUM, Malaysia

arifullahms88@gmail.com
sidra.pervez@hitecuni.edu.pk
hn230030@student.uthm.edu.my
alinashah597@gmail.com
muazam.ali@hitecuni.edu.pk

Abstract: This study investigates the detection of deepfake images and videos on social media platforms such as Instagram for forensic analysis using hybrid-learning approaches. It highlights the critical importance of safeguarding privacy and authenticity in digital media. The background draws attention to the growing threat posed by deepfakes, which pose significant challenges across multiple domains, such as politics and entertainment. Existing methods often depend on visual features specific to a dataset and struggle to generalize across different manipulation techniques. Moreover, most approaches focus exclusively on either temporal or spatial features, which limits their capacity to identify complex anomalies involving fused facial features like the mouth, nose and eyes. Important solutions to these challenges include Convolutional Neural Network (CNN), Recurrent Neural Networks (RNN) and hybrid architectures that simultaneously capture spatial and temporal information in deepfake content, such as Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM), Gated Recurrent Unit (GRU) and Vision Transformers (ViT). Additionally, this paper introduces a novel combination of artifact inspection and facial landmark recognition to enhance detection accuracy and employs Gated Recurrent Units (GRUs) and Vision Transformers (ViT) for data augmentation thereby improving model robustness. The effectiveness of the proposed approach is validated through experiments demonstrating substantially improved deduction accuracy, with improvement exceeding 1.5% across multiple datasets. However, several challenges remain, including limited robustness to noise, difficulty in detecting deepfakes in compressed video formats, and dataset imbalances issues. The proposed enhanced hybrid model exhibits superior detection performance while maintaining adaptability across multiple datasets. Future research will focus strengthening model generalization to effectively counter emerging deepfake generation techniques.

Keyword: Deepfakes, Image forensics, Privacy protection, Hybrid model, Web technology

1. Introduction

Information is now widely accessible due to the development of web technology. Every day, a great deal of information is transmitted through print and online media, yet it can be challenging to assess its veracity [1]. The deliberate attempt to damage or enhance the reputation of a company, institution, or individual in order to gain financial or political advantages is known as the dissemination of false information [2]. A wide range of fabricated stories, falsified news articles, and digitally manipulated images circulate on social media. The consumption and distribution of information, as well as various modes of communication including blogging and texting, have been completely transformed by digital platforms [3]. Deepfake technology benefits the fashion industry by allowing customers to make fashion decisions more quickly. Furthermore, this technology has the potential to help individuals with Alzheimer's disease communicate by enabling them to engage with a digitally modified version of their younger selves [4]. Beyond detection, forensic analysis in the deepfake arena covers serval important study areas. These include techniques designed to function effectively in authentic, real-world situations, as well as deepfake attribution, recognition, and passive and active authentication. Deepfake attribution and recognition seek to track the origins of synthetic content by identifying the precise models utilized in its production. This entails examining the traces left by generative models in order to attribute content to its source. Without the need for additional embedded data, passive authentication methods concentrate on assessing the legitimacy of media using its natural characteristics, such as statistical anomalies. For retrospective analysis, these techniques work especially well [5]. Conversely, active authentication methods incorporate verifiable data, such as digital watermarks or cryptographic signatures, into media during the creation process. Deepfake technology is developing very quickly. Apart from the conventional ways of modifying particular regions of an image or video, computer graphics and deep learning (DL) have introduced in a number of new approaches [6]. It has been demonstrated that auto-encoders and Generative Adversarial Networks (GANs) aid in producing high-quality face synthesis with a high degree of photorealism. Additionally, a segmentation map

can be used to create fake videos or image. A written representation or a drawing can be used to create an image. Similarly, a series of auditory stimuli can change a person's facial features. Often, the modification involves altering pre-existing photos or videos [7].

Deepfake Generation Process: A class of machine learning techniques known as generative models is designed to discover the underlying distribution of a dataset and produce new samples that closely resemble the original data. Generative models seek to capture the data distribution $pdata$ and create new samples with comparable features, in contrast to discriminative models, which concentrate on distinguishing between classes or generating predictions [8]. These models have gained significant attention in a variety of fields, such as data augmentation, text generation, music composition, and image synthesis. Generative Adversarial Networks and Diffusion Models are two most widely used generative modeling techniques. Although the goal of both approaches is to create high-quality synthetic data, their underlying principles and learning paradigms are very different. Generating a deepfake is a multi-step pipeline that uses cutting-edge machine learning algorithms to create incredibly lifelike media [9].

Deepfake Detection: As deepfake, technology develops and creates more realistic synthetic media, ensuring the authenticity of digital information has become a crucial challenge in multimedia forensics. In addition to providing the foundation for forensic analysis, attribution, and authentication, detecting modified content is essential for mitigating the risks of disinformation, identity fraud, and threat to media integrity concerns [10]. This section offers a structured summary of detection methodologies, ranging from traditional forensic methods to contemporary data-driven strategies. The primary goal of early forensic techniques was to detect obvious spatial anomalies, such as compression artifacts, pixel-level irregularities, and inconsistent illumination [11]. Although these techniques remain relevant, they frequently fail to account for highly sophisticated adjustments. To overcome these constraints, deep learning algorithms utilize data-driven patterns to identify subtle discrepancies that are invisible to the naked eye [12].

2. Related Work

Table 1 below presents a summary of the related work, showing the authors' names, research areas, publication years, and findings.

Table 1: Summary of related work

Ref	Technique	Findings
[13]	Closest Neighbor Classification using pretrained vision-language model	Conventional DNN classifiers are unable to identify fake images from cutting-edge generative models. The closest neighbor classification approach increased accuracy by 25.90% and mAP by 15.07 m on unobserved diffusion and autoregressive models.
[14]	Statistical Consistency Assault (StatAttack) and MStatAttack	Deepfake detecting statistical discrepancies are decreased using MStatAttack. The updated MStatAttack added layers of deterioration and adjusted combination weights. It may be able to get around sophisticated Deepfake detection systems because it operated in both white-box and black-box scenarios spanning detectors and datasets.
[15]	Hybrid technique combining Residual Network and KK-NN methods	Used a methodical strategy to classify Deepfake pictures with an accuracy of 89.5%. The hybrid approach was strong and successful in identifying Deepfake, indicating that it may lessen the dangers of propaganda and defamation on social media like Instagram.
[16]	ID-unaware Deepfake Detection Model	Enhanced generalization of binary classifiers because of implicit identity leakage. To address this issue, the ID-unaware model was proposed, and it performed better than previous methods over a wide range of datasets.
[17]	Deepfake Predictor (DFP) using VGG16 and CNN architecture	Created the DFP framework, which had a 94% accuracy and 95% precision rate in identifying Deepfake. The technique proved more effective at identifying Deepfake media and resolving ethical issues than transfer learning and other contemporary techniques.

3. Proposed Model

The proposed methodology addresses deepfake detection and privacy protection on social media platforms such as Instagram in image forensics by combining Gated Recurrent Units (GRUs) with Vision Transformers (ViTs). Sequential dependencies and temporal artifacts found in image and video sequences are captured by GRUs. High-dimensional global spatial characteristics and structural patterns are extracted from image patches using ViTs. The model improves sensitivity to both spatial and temporal inconsistencies typical of deepfakes by fusing

the ViT's proficiency in contextual image representation with the GRU's capacity to model temporal changes. The framework uses GRU to identify minute temporal irregularities in motion coherence and facial expressions. ViT meanwhile assist by identifying frequency domain distortions, visual artifacts, and inconsistencies unique to GAN-generated content. Strong feature fusion from both sequential and patch-based perspectives is ensured by this hybridization. Figure 1 presents the proposed model structure along with its different components.

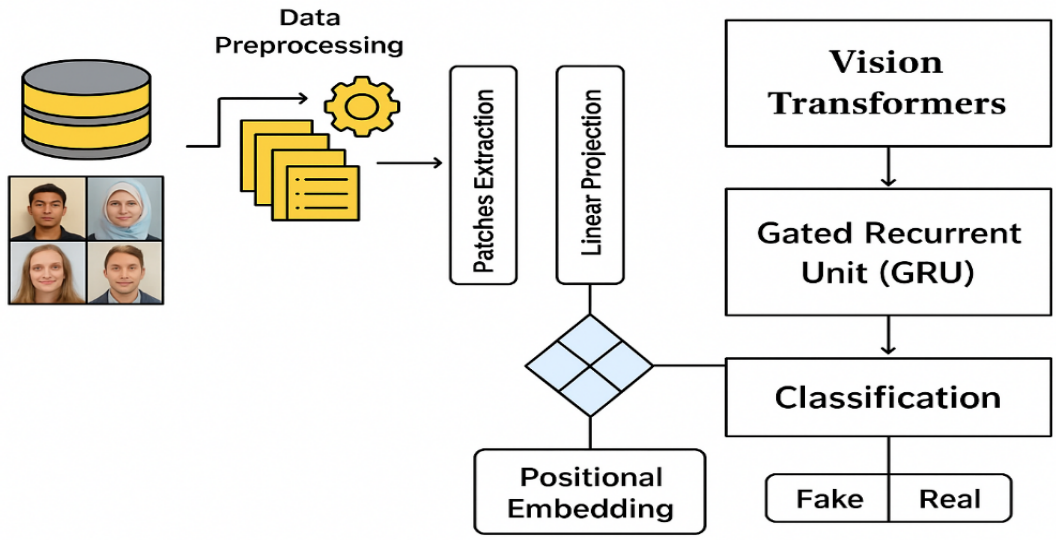


Figure.1: Proposed model for deepfakes detection in social media like Instagram

An outline of our deepfake detection procedure is presented in Figure 1, with particular attention paid to steps such as data preprocessing, model training, and binary classification topology categorization. When paired with behavioral or biological signal cues, this integration facilitates multi-modal forensic analysis. By limiting the availability of raw data while preserving high detection accuracy, the approach prioritizes privacy. Overall, this GRU–ViT hybrid framework provides a robust, scalable, and explainable defense against deepfake threats in digital forensics. A brief mathematical explanation of the proposed hybrid model, which combines privacy-aware training (federated learning with differential privacy) and image forensics (deepfake detection and localization) using a Vision Transformer (ViT) backbone and a Gated Recurrent Unit (GRU) module, is provided below. This section introduces the notation, privacy-aware federated update equations, model heads, loss functions, core equations (1), and the feature fusion process described in Equation (1).

$$\ell - 1 = LN(zi\ell - 1)MHA(\ell - 1)\sum = 1head(\ell - 1)MHA(z{\sim}\ell - 1)i = m = 1\sum Hheadm(z{\sim}\ell - 1)i\ell = \ell - 1 + MHA(\ell - 1) + MLP(LN(\ell - 1))zi\ell = zi\ell - 1 + MHA(z{\sim}\ell - 1)i + MLP(LN(zi\ell - 1)) \quad \textbf{(1)}$$

The Transformer model's layer operations are described in the four lines that follow. Prior to being sent into a Multi-Head Attention (MHA) mechanism, where attention is computed over several heads and their results are averaged, the input from the preceding layer, l − 1 l−1, is first normalized using Layer Normalization (LN). Updated contextualized representations of the input are thus produced. The original input zl−1 z l−1 is then added to the attention output by a residual connection, and the features are further refined by a Layer Normalization and Multi-Layer Perceptron (MLP). Ultimately, the residual input, the MHA output, and the MLP transformation are combined to create the output zlz l at layer ll, guaranteeing stability, richer representation learning, and improved gradient flow across the deep network.

$$t = Er(xt) = (+h - 1+)zt = \sigma(Wzrt + Uzht - 1 + bz)'(+h - 1+)rt' = \sigma(Wrrt + Urht - 1 + br)h{\sim} = tanh\,(h + h(' \odot h - 1) + h)h{\sim}t = tanh(Whrt + Uh(rt' \odot ht - 1) + bh)h = (1-) \odot h - 1 + \odot h{\sim}ht = (1 - zt) \odot ht - 1 + zt \odot h{\sim}t \quad \textbf{(2)}$$

The updating procedure in a Gated Recurrent Unit (GRU) is described by in equations (2) Initially, the update gate z t and reset the gate rtr t are calculated using sigmoid functions, which determine the amount of historical data to retain or discard. Next, a potential concealed state h ~ t h ~ t is produced with the reset gate's assistance, and the ultimate hidden state ht h t is calculated as the weighted sum of the prior state ht−1 h t−1 and the candidate state, which the update gate regulates.

4. Datasets

A sizable dataset called WildFake was created to facilitate the deduction of AI-generated images. It includes a wide range of synthetic photos from various generative models and open communities, ensuring a wide range of styles and content. Images in the dataset are categorized hierarchically by generative model types, architectures, weights, versions, and release dates in a hierarchical manner [17]. Both authentic and synthetic facial photos are included in the sample. We employed a validation set of 76,161 photos for our assessment, of which 38,080 were authentic and 38,081 were fraudulent. The evaluation metrics used include accuracy, precision, recall, and the F1-score.

5. Experiment Results and Discussion

This study evaluates both standard GU models and the ViT-based model for deepfake detection.

The results demonstrate a notable performance advantage for the ViT, which is attributed to its ability to integrate both local and global features through self-attention mechanisms. A thorough presentation of the results is provided in the following section and illustrated in Figure 2.

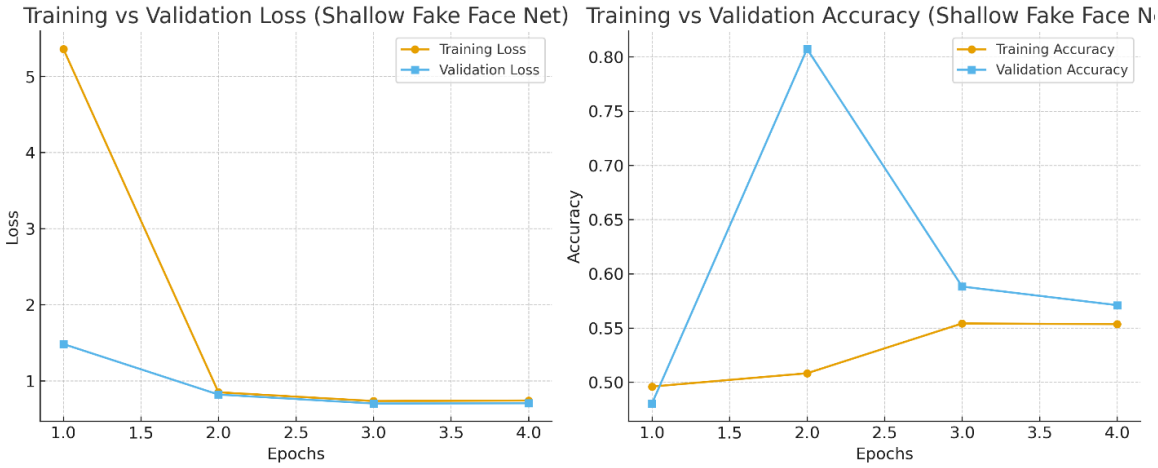


Figure 2: Accuracy and Loss values over training dataset

Figure 2 displays the accuracy and loss of curves for both training and validation. The Y-axis in training and validation accuracy displays accuracy, which ranges from 0 to 1 and indicates how well the model predicts. Complete runs of the training dataset, or epochs, are the units on the X-axis. The training accuracy is displayed by the yellow line (R) graph, which begins approximately 0.4962 and gradually increases to almost 0.5538 in the fifth epoch. The upward trend indicates that the model is getting better at what it does by learning from the training data. The validation accuracy is indicated by the orange line, which starts at 0.4803, experiences a slight decline after the second epoch, and subsequently increases to 0.5710 by the fifth epoch. Two high performing models for identifying GAN-generated fused artifacts in facial features TCN and CNN-LSTM with GAN. For strong generalization, we achieve excellent alignment between the accuracy and loss curves of both models. CNN and CNN-GRU both provide dependable stability, however CNN-GRU does not achieve the same accuracy levels. CNN performs reasonably well, but it could use some further enhancements to handle the intricacy of fused GAN-generated features. Our findings build on the research of [18], where the authors employed CNN models like VGG16, VGG19, and the InceptionV3 model to detect deepfakes with an accuracy of up to 90%. The effectiveness of attention-based architectures in deepfake identification is demonstrated by this work, which uses PViT to provide a 9% performance boost with 99% accuracy. As indicated in Figure 3, the outcomes of both investigations are comparable because they use the same Kaggle dataset. The fact that ViT performs so well demonstrates the benefit of processing image patches in their holistic form, which makes it possible to identify changes across local and global inputs. This approach addresses several limitations of CNN architectures and reinforces the importance of transformer models in multimedia forensics.

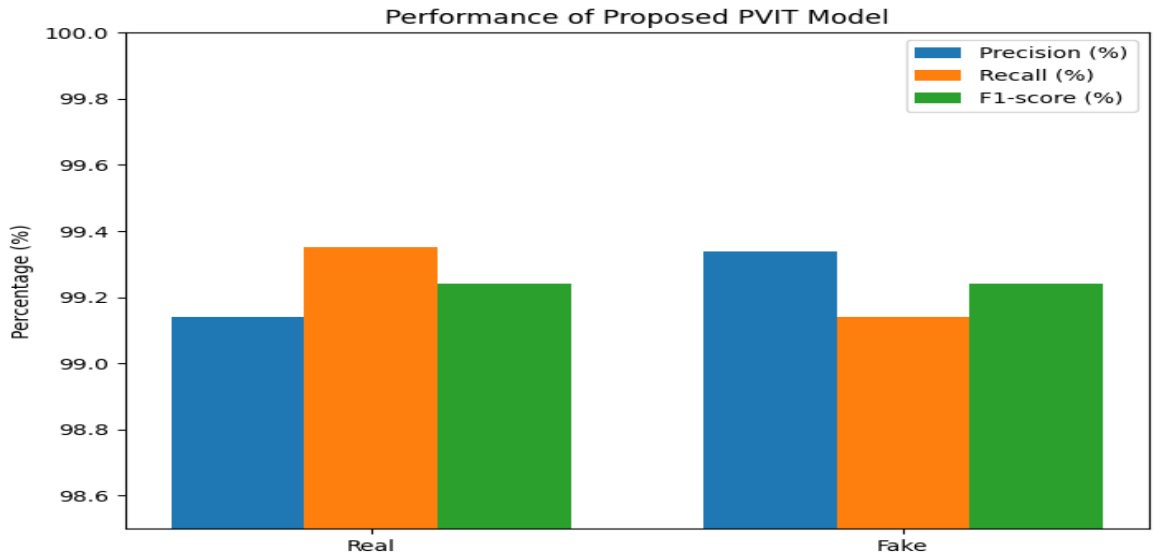


Figure 3: Performance of Proposed PVIT Model

According to these findings, the ACC-GAN model has improved the mean values of performance metrics like 95.7% PRE, 99.3% SEN, 90.0% SPE, and 98.40% ACC. Additionally, the suggested ACC-GAN model attains a ρ-value of less than 5% (1.6%), indicating that the outcomes obtained by ACC-GAN are noteworthy. With 1.4% accuracy, 3.3% sensitivity, 1.6% specificity, 3.8% precision, and an 8.4% ρ-value, the suggested model outperforms GPT-4o. The suggested model achieves minimum SD values of 0.7% accuracy, 0.3% sensitivity, 2.6% specificity, 2.8% precision, and 0.8% ρ-value when taking into account the SD value of the evaluation metrics.

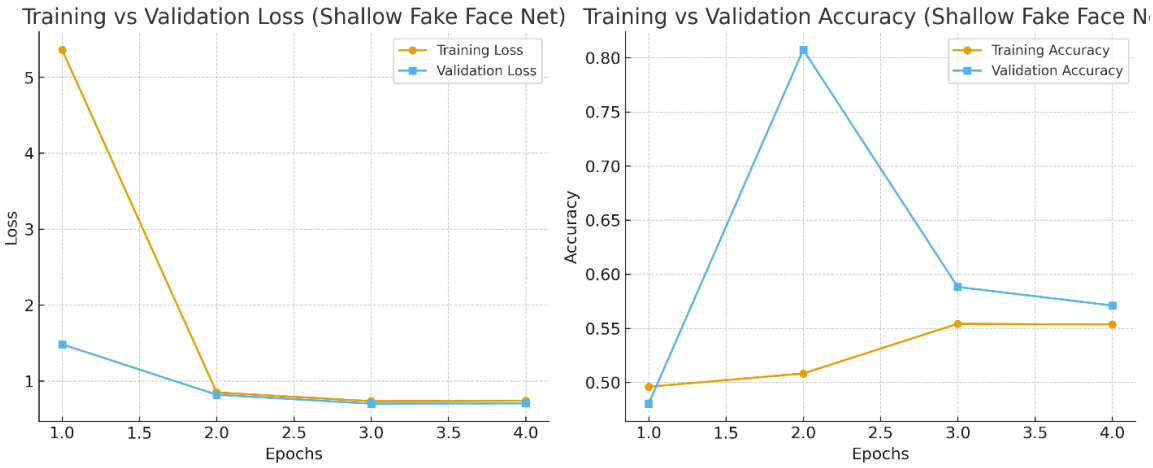


Figure 4: Accuracy and Loss values of CNN-PCA approach

Figure 4 illustrates the training and validation curves for the CNN-Principal Component Analysis approach, focusing on accuracy and loss. Accuracy, a number between 0 and 1, is displayed on the Y-axis in training and confirmation accuracy, indicating how well the model predicted. The X-axis displays the epochs, which indicate the number of full runs of the training dataset. The training accuracy is displayed by the yellow line. It begins at roughly 0.5138 and increases to roughly 0.7526 by the fifth phase, demonstrating how the model continues to get better as it absorbs information from the training set. The validation accuracy is displayed by the orange line. After the first epoch, it begins at roughly 0.5908, falls to roughly 0.5123, and then increases to roughly 0.5833. Based on these results, the proposed model achieves a 1.5% performance improvement across multiple datasets using various parameters.

6. Conclusion

Numerous scholars have been working extensively to identify misleading information using a variety of computational approaches. This review study demonstrates existing methods for the identification and classification of deepfake images. Because deepfake images are so realistic, it might be difficult for conventional

image analysis techniques struggle to emphasizes on the difficulty in determining the authenticity of the images and videos. There are various techniques for creating deepfakes, from deep neural networks to GANs. The model needs to be resistant to different methods of producing deepfakes. To ensure its efficacy over time, the solution should be designed to generalize and adapt effectively to emerging deepfake generation technologies. Adversarial attacks can be used by malicious actors to covertly alter deepfake content in order to fool detection systems. The legitimacy and security of digital content are seriously threatened by the growing realism and accessibility of modified images, particularly those produced by sophisticated AI methods like GANs. Even though they are effective and interpretable, traditional picture forensic techniques are unable to combat the sophistication of contemporary visual forgeries. Convolutional Neural Networks (CNNs), in particular, are deep learning-based models that provide high detection accuracy but have limited transparency and are susceptible to hostile attacks. This study addresses these constraints by proposing a hybrid strategy that combines deep learning methods with traditional forensic analysis. The proposed model achieves high accuracy, better generalization, and increased interpretability by fusing CNNs' feature-learning capabilities with Photo-Response Non-Uniformity analysis, compression artifacts, and other manually created features. Experimental results confirm the effectiveness of this hybrid framework, demonstrating that it outperforms standalone methods in identifying both traditional and AI-generated manipulations. Additionally, the model is more resilient to adversarial environments and offers clearer justifications for its conclusions, which makes it more appropriate for practical application in high-stakes fields like digital security, legal investigations, and journalism. Future research should examine additional fusion mechanism optimization, expand the system to manage video-based alterations involving videos, and enhance real-time detection capabilities. Such reliable and comprehensible detection techniques are essential for preserving confidence in visual media in the digital age, where disinformation is only becoming more widespread and complicated.

Ethics Declaration: This research did not require ethical approval as it did not involve any secret human participants, personal data collection, or sensitive information. All the sources and examples cited in this paper are publicly available and properly referenced/cited.

AI Declaration: AI tools, such as Google Gemmini and DeepSeek, were used to assist in refining academic language, and organizing the structure of the manuscript. All content generated by AI was critically reviewed, edited, and verified by the author(s) to ensure accuracy, originality, and adherence to scholarly standards. However, this manuscript similarity index was checked using plagiarism detection software and found to be within the acceptable range, indicating proper academic integrity and originality.

References

Camacho, I. C., et al. (2021). A Comprehensive Review of Deep-Learning-Based Image Forensics Techniques. *PMC*.

Gandhi, A., & Jain, S. (2020). Adversarial Perturbations Fool Deepfake Detectors.

Gandhi, A., & Jain, S. (2020). Adversarial Perturbations Fool Deepfake Detectors.

Hong, Y., Feng, J., Chen, H., Lan, J., Zhu, H., Wang, W., & Zhang, J. (2025, April). Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 4, pp. 3500-3508).

Killekar, A., et al. (2020). Two-Branch Recurrent Network for Isolating Deepfakes in Videos.

Kim, M., et al. (2021). FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning.

Kim, M., et al. (2021). FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning.

Kumar, B. A. (2025). Hybrid CMNV2: DeepFake Faces Classification and Detection. *ScienceDirect*.

Kumar, S., & Narang, R. (2025). Combating Digitally Altered Images: Deepfake Detection Using Vision Transformer.

Qureshi, S. M., et al. (2024). Deepfake Forensics: A Survey of Digital Forensic Methods for Deepfake Detection. *PMC*.

Singh, S. (2025). Unmasking Digital Deceptions: An Integrative Review of Deepfake Generation and Detection. *ScienceDirect*.

Sohail, S., Sajjad, S. M., Zafar, A., Iqbal, Z., Muhammad, Z., & Kazim, M. (2025). Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning. *Information*, 16(4), 270.

Verdoliva, L. (2021). Multimedia Forensics: Challenges and Opportunities in the Age of Deepfakes. *IEEE*.

Verdoliva, L. (2021). Multimedia Forensics: Challenges and Opportunities in the Age of Deepfakes. *IEEE*.

Wang, R., et al. (2020). FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking. Shan, S., et al. (2020). Fawkes: Protecting Privacy against Unauthorized Deep Learning Models.

Wang, R., et al. (2020). FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking.

Yadav, S. (2025). Enhancing Fake Image Detection with a Hybrid Approach. *IJSAT*.

Zhang, Y., et al. (2025). A Review of Generated Images and Deepfake Detection Techniques. *Springer*.

VessiGuard: AI-Driven Anomaly Detection for Maritime Cyber Defence

Ravi Varma Kanumuri, Laavanya Rachakonda, Hosam Alamleh, Bilge Karabacak and Ulku Clark Yaylacicegi
University of North Carolina, Wilmington, USA

rk1260@uncw.edu
rachakondal@uncw.edu
alamlehh@uncw.edu
karabacakb@uncw.edu
clarku@uncw.edu

Abstract: The maritime industry is undergoing a rapid digital transformation, driven by the adoption of technologies such as the Automatic Identification System (AIS), advanced navigation software, and onboard Internet of Things (IoT) sensors. These innovations have significantly improved operational efficiency, safety, and situational awareness. However, this increasing reliance on interconnected digital systems also expands the sector's exposure to cyber threats. Traditional rule-based monitoring and siloed intrusion detection systems often fail to identify coordinated multi-modal attacks, leaving vessels vulnerable to sophisticated, stealthy intrusions. The dual nature of this transformation underscores the urgent need for more sophisticated and adaptive cybersecurity strategies. This study introduces VessiGuard, an AI-driven anomaly detection system designed to detect and mitigate abnormal vessel behaviour as an early indicator of potential cyber intrusions, system failures, or operational anomalies. The approach leverages two complementary artificial intelligence techniques: Long Short-Term Memory (LSTM) neural networks, which model temporal dependencies in vessel movement, and Isolation Forest algorithms, which excel at detecting rare and unusual behaviour patterns. By fusing navigational telemetry with operational technology (OT) sensor readings, specifically engine temperature and fuel consumption, the model creates a unified, cross-domain anomaly score that is robust against single-variable manipulation. A prototype anomaly detection system was implemented and evaluated using controlled simulation and publicly available maritime datasets that reflect real-world operational scenarios. Results demonstrate that VessiGuard effectively detects anomalies, including GPS spoofing, sensor drift, and structured interference. Experimental validation indicates a detection accuracy of approximately 94.2% for trajectory anomalies and 92.8% for sensor deviations. Furthermore, the system demonstrates modality-specific responsiveness, identifying operational sensor faults in under four minutes while accurately accumulating evidence for trajectory deviations within five to eight minutes. This work presents a practical pathway towards adaptive, data-driven cybersecurity solutions by situating anomaly detection within the broader maritime operational ecosystem. The findings highlight how AI-based anomaly detection can complement existing maritime defence mechanisms, support decision-making under dynamic threat conditions, and improve incident response readiness. Furthermore, the results lay the groundwork for future research into autonomous and semi-autonomous detection architectures, ultimately contributing to a more resilient, secure, and intelligence-driven digital maritime domain.

Keywords: Maritime cybersecurity, Anomaly detection, Machine learning, Cyber-Physical systems, Cyber resilience, AIS security

1. Introduction

The maritime industry underpins global trade, with over 90% of goods transported by sea. As shipping operations evolve, vessels increasingly rely on digital technologies, such as the Automatic Identification System (AIS), GPS navigation, onboard Internet of Things (IoT) sensors, and advanced communication networks, to enhance operational efficiency, safety, and situational awareness (Kim and Joe, 2024). However, this digital transformation has also introduced complex cybersecurity risks. Modern ships now function as floating cyber-physical systems, and their growing connectivity exposes them to threats such as GPS spoofing, data manipulation, unauthorised access, and coordinated cyberattacks on operational technology (OT) systems. These vulnerabilities can disrupt navigation, endanger crew safety, and jeopardise global supply chains (Bobrovnikova et al., 2022; School, 2022).

Detecting such threats remains challenging. Unlike mechanical failures, cyberattacks are often stealthy, adaptive, and embedded within legitimate data streams. Conventional rule-based systems struggle to detect these subtle anomalies in real time, often generating excessive false alarms or missing sophisticated attacks. As maritime systems become more interconnected (IMO, 2017a; IMO, 2017b; BIMCO et al., (2020), there is an urgent need for intelligent, automated, and adaptive cybersecurity solutions capable of context-aware, early detection of abnormal behaviours (Maganaris et al., 2024; Shit et al., 2025).

This paper introduces VessiGuard, an AI-driven anomaly detection system designed to enhance maritime resilience through intelligent anomaly detection. It employs advanced artificial intelligence (AI) techniques, including Long Short-Term Memory (LSTM) neural networks and Isolation Forest algorithms, to analyse AIS, GPS, and sensor data for deviations from expected operational patterns. These models detect subtle temporal anomalies and rare behavioural outliers as indicators of cyber intrusions, system malfunctions, or unauthorised activities.

This work focuses on the implementation and evaluation of its core anomaly detection engine. A prototype system, validated using simulated maritime datasets, demonstrates effective detection of anomalies such as GPS spoofing, sensor manipulation, and unauthorised access attempts.

By combining state-of-the-art AI with a modular system design, VessiGuard represents a significant step towards autonomous, adaptive, and data-driven maritime cybersecurity. Beyond real-time anomaly detection, it lays the groundwork for future integration of resilient response mechanisms, contributing to safer and more secure maritime operations in an increasingly digital era.

2. Related Work

The maritime industry's increasing digitalisation has prompted significant research into cybersecurity, particularly in anomaly detection and secure data handling. Existing studies have explored diverse approaches ranging from trajectory-based monitoring and sensor analysis to intrusion detection systems. This section reviews these developments, highlighting their strengths and limitations, and identifies the gap the proposed VessiGuard system aims to address.

AIS-based Vessel Behaviour Analysis: AIS data forms the backbone of most maritime anomaly detection systems. Traditional approaches rely on rule-based mechanisms such as geofencing and speed thresholds to flag deviations from expected vessel behaviour. While these methods are straightforward, they often suffer from high false-positive rates and limited adaptability (Riveiro et al, 2018). Machine learning methods have enhanced AIS-based monitoring by learning standard movement patterns, using models like Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to capture temporal dependencies. Despite their effectiveness, these models are often data-hungry and require high-quality labelled datasets, which are scarce in the maritime domain (Davari et al, 2021; Ribeiro et al, 2023). Additionally, most AIS-focused studies operate exclusively on navigational trajectories and do not integrate physical sensor states, leaving them vulnerable to multi-modal cyber-physical spoofing.

*Sensor and Engine Anomaly Detection***:** Beyond navigation data, several studies have examined anomaly detection using sensor and machinery data. Deep learning models such as convolutional neural networks (CNNs) and autoencoders have been used to monitor operational parameters, while Transformer-based models identify complex, non-linear patterns and deviations from normal operating conditions (Thakur et al, 2025). Such methods effectively detect cyber-induced manipulation or faults but face deployment challenges due to sensor variability across vessels and the need for domain-specific feature engineering (Amro et al, 2022). Moreover, many approaches focus solely on detecting anomalies without integrating incident response mechanisms or cross-referencing them with navigation data.

Intrusion Detection Systems in Maritime Networks: The convergence of information technology (IT) and operational technology (OT) aboard vessels introduces new network-based vulnerabilities. Intrusion detection systems (IDS) have been developed to identify malicious traffic, unauthorised access, or abnormal communication patterns (Wang et al, 2022; Brenner et al, 2023). Unsupervised learning techniques such as clustering and one-class classification have been applied to network traffic data to detect outliers without requiring extensive labelled datasets (Talukder et al, 2024; Nguyen et al, 2020). However, most IDS approaches are limited in their ability to correlate network anomalies with physical system states, offering little insight into the physical impact of cyber events.

2.1 Summary and Research Gap

Although substantial progress has been made in maritime cybersecurity, most approaches target isolated aspects, such as AIS-based anomaly detection, sensor monitoring, or network intrusion detection, rather than providing a unified defence solution. These siloed approaches result in fragmented situational awareness, which is insufficient for detecting coordinated or multi-modal cyberattacks. Furthermore, few frameworks combine real-time detection with secure logging and automated incident response; a brief comparison is mentioned in Table 1.

Table 1: Brief comparison with the existing State-of-the-Art Literature

Research Area	Data Type	Methods Used	Strengths	Limitations	Gap Addressed by VessiGuard
Rule-based AIS anomaly detection Riveiro et al. (2018)	AIS trajectories	Geofencing, speed rules	Simple, interpretable	High false positives; no temporal modeling	ML-based sequence learning
Deep learning AIS models (RNN/LSTM) Davari and Aguiar (2021)	AIS/GPS	Sequential neural networks	Captures temporal dependencies	Requires high-quality labeled data; ignores sensors	fuses navigation + sensor behaviors
Sensor anomaly detection Nguyen et al. (2020)	Engine, fuel, compass	Autoencoders, CNNs, Transformers	Detects subtle OT faults	Vessel-specific; noisy logs; no link to navigation	unites OT + navigation streams
Maritime IDS Amro et al. (2022)	Network packets	One-class, clustering, MTL	Detects malicious traffic	No physical context; blind to sensor/trajectory attacks	cyber-physical anomaly detection
AIS deep-generative models Wang et al. (2022)	AIS	Probabilistic trajectory modeling	Strong spatial-temporal modeling	Cannot detect OT manipulation	integrates OT and GPS/compass features

VessiGuard addresses these gaps by integrating AI-based anomaly detection, cross-domain data fusion, and laying the foundation for integrity mechanisms into a cohesive, modular architecture. This approach enhances detection accuracy, ensures secure event traceability is possible, and lays the foundation for adaptive, automated maritime cybersecurity systems.

3. Proposed Framework

3.1 System Overview and Design Rationale

VessiGuard is designed as a modular anomaly detection system that fuses navigation, telemetry, and operational sensor data using machine-learning techniques. Rather than serving as a full cybersecurity framework, VessiGuard focuses specifically on detecting anomalous patterns that may indicate spoofing, sensor manipulation, or cyber-physical irregularities. The system employs a hybrid architecture that combines sequential and statistical learning to enhance robustness across various types of anomalies.

The architecture is built around four core layers: data acquisition, anomaly detection, score fusion, and a conceptual human-centered interface. The current implementation primarily focuses on the data acquisition and anomaly detection layers. The design rationale rests on two core principles:

- Cross-modal Integration: A unified system that analyzes AIS/GPS signals alongside OT sensor data improves detection accuracy against cyber incidents that frequently affect multiple data sources simultaneously.
- Complementary ML Specialisation: LSTM autoencoders specialise in capturing temporal vessel behaviour, while Isolation Forest models detect abrupt deviations in sensor data. A fused approach mitigates weaknesses inherent in using either model independently.

3.2 Data Acquisition Layer

The foundation of VessiGuard is a data ingestion pipeline that aggregates multi-modal maritime telemetry from diverse onboard sources. This includes AIS broadcasts, GPS navigation data, compass readings, engine and fuel sensor metrics, and other operational parameters essential to vessel performance. The system operates with simulated datasets and real-time feeds, ensuring compatibility with maritime communication standards, such as NMEA 0183 and NMEA 2000 (Kim, H., & Joe, I., 2024). This layer performs data cleaning, normalisation, resampling, and temporal alignment to ensure consistency before feeding into the ML models. The

preprocessing step is essential for reducing false alarms and improving the reliability of anomaly detection by ensuring that the downstream ML models receive stable, synchronised inputs.

3.3 AI-Driven Anomaly Detection Engine

The core of VessiGuard's current implementation is the Anomaly Detection Engine, which comprises two distinct, complementary models:

1. Long Short-Term Memory (LSTM) Autoencoder: This model learns temporal dependencies in vessel behaviour, capturing patterns in continuous signals like vessel movement or sensor readings. It reconstructs expected trajectories; deviations in the reconstruction error indicate sequence-level anomalies (Hochreiter and Schmidhuber, 1997). This model detects temporal drifts or trajectory inconsistencies.
2. Isolation Forest (IF) Algorithm: This lightweight, unsupervised model excels at detecting rare, non-sequential deviations that differ statistically from the majority of the data (Liu, Ting and Zhou, 2008). IF is particularly effective at identifying subtle irregularities in sensor or telemetry data that may signal unauthorised access or latent system faults. It detects statistically rare events in static or low-dimensional feature spaces.

Together, these models form a hybrid detection pipeline. The LSTM autoencoder learns vessel motion patterns over time, while the Isolation Forest detects instantaneous deviations in sensor readings, providing complementary strengths. This dual-model strategy ensures resilience against both gradual and abrupt irregularities in maritime datasets.

4. Implementation, Methodology, and Experimental Setup

4.1 Research Question and Objectives

This study investigates the following research question: How can machine learning algorithms, specifically LSTM and Isolation Forest models applied to AIS, GPS, and OT data, enhance real-time anomaly detection in maritime cybersecurity? To address this, the primary objective is to develop and evaluate a prototype system capable of identifying operational and navigational anomalies, indicating potential cyber intrusions or system failures, and providing actionable intelligence to vessel operators before incidents escalate.

4.2 Scope of the Study

Four core areas define the scope of this work:

- Prototype Development: Designing a functional system to monitor AIS, GPS, and simulated OT telemetry.
- Model Application and Real-Time Detection: Applying LSTM for sequential behaviour modelling and Isolation Forest for multi-dimensional outlier detection.
- Simulation-Based Evaluation: Validating detection accuracy, false alarm rates, and responsiveness using simulated datasets.
- Integration Considerations: Ensuring compatibility with existing vessel monitoring workflows without significant infrastructure changes.

4.3 Project Motivation and Contribution

VessiGuard addresses the gap in deployable maritime security solutions by presenting a practical, lightweight anomaly detection system. Its contributions are threefold: first, it introduces a novel **hybrid ML approach** that combines LSTM-based sequence reconstruction with Isolation Forest outlier detection for multimodal vessel data. Second, the system features a **real-time, edge-ready design** suitable for deployment on vessel-based computing hardware. Finally, it establishes an **operational integration pathway**, demonstrating how advanced anomaly detection can be embedded within broader cybersecurity architectures.

4.4 Dataset, Preprocessing, and Feature Engineering

The dataset comprises Automatic Identification System (AIS) and Global Positioning System (GPS) telemetry (latitude, longitude, Speed Over Ground (SOG), Course Over Ground (COG)) alongside simulated Operational Technology (OT) data, including engine temperature and fuel rate (Varma, 2024). Data were resampled to a 1-minute cadence and filtered to remove invalid MMSI, zero coordinates, or implausible speeds (> 50 kn). Preprocessing involved deduplication, chronological sorting, and interpolating short gaps (≤ 5 min) along great-

circle paths, while longer gaps initiated new segments. GPS coordinates (WGS84) were converted to a local east–north–up (ENU) frame to **stabilise** kinematic computations. Feature engineering involved computing kinematic metrics (displacements Δx/Δy, acceleration, jerk, turn rate, and cross-track error) and geofencing parameters (distance to coastline/centreline and harbour proximity) over a sliding window of T. Behavioural indicators included stop/start flags, loitering scores, and sudden heading flips. OT features comprised z-scored engine temperature and fuel rate residuals against a kinematic load model, alongside window aggregates (means, variances, and stop fractions).

4.5 Methodology

The VessiGuard prototype models vessel behaviour using sliding windows of telemetry data (T x F) derived from AIS, GPS, and simulated OT sensors. The system computes an anomaly score s(X) ∈ [0,1], where higher values indicate greater deviation, targeting real-time inference on edge hardware. The framework integrates two complementary models. First, a **Sequence Model** employing a two-layer LSTM autoencoder (hidden size 128, dropout 0.2) reconstructs sequential inputs, using reconstruction error as the anomaly signal. To support transferability across fleets with limited data, bottleneck adapters are inserted between layers, allowing fine-tuning of only ~5% of parameters. The loss function combines feature-weighted mean squared error with a temporal smoothness penalty:

$$\mathcal{L} = \sum_{t=1}^{T} w \, \|\hat{x}_t - x_t\|_2^2 \; + \; \lambda \sum_{t=2}^{T} \|(\hat{x}_t - \hat{x}_{t-1}) - (x_t - x_{t-1})\|_2^2$$

Window-level errors are **normalised** to produce the score s_{LSTM}, with per-timestep scores providing interpretability.

In parallel, a **Tabular Outlier Model** using an Isolation Forest (n_estimators=300, contamination=0.01–0.10) processes aggregated features to detect abrupt sensor deviations, such as fuel spikes or temperature drifts. Its outputs are scaled to s_{IF} ∈ [0,1]. Finally, these outputs are fused to provide a robust anomaly score.

$$s = \alpha * s_LSTM + (1 - \alpha) * s_IF, \alpha \in [0.3,0.7]$$

Alerts are triggered when s≥ τ, with the threshold τ selected to **maximise** the F1 score or maintain a target false alarm rate.

4.6 Simulation and Validation Setup

Validation was conducted entirely in a controlled simulation environment using injected anomalies such as:

- *GPS Spoofing*: Artificial trajectory deviations from the vessel trajectory were injected to simulate navigation signal tampering (Varma, 2024).
- *Sensor Drift*: Gradual manipulation of OT readings through artificial spikes and drops
- *Compass Perturbations*: Sudden heading flips to mimic compass spoofing
- *Structured Noise*: Periodic false signals to evaluate how well the models handled structured noise.

Cross-validation was performed on time-split datasets to avoid leakage of future information into training. In addition, region-holdout experiments were conducted, where one port's data was excluded from training to test generalisation. The Isolation Forest and LSTM-AE models were validated independently and in fused mode to assess improvements from score combination.

5. Results and Discussion

5.1 Simulation Scenarios and Evaluation Metrics

Validation was conducted in a controlled simulation environment using injected anomalies to test system resilience: **GPS Spoofing** (trajectory deviations), **Sensor Drift** (gradual OT manipulation), **Compass Perturbations** (sudden heading flips), and **Structured Noise**. Performance was evaluated using **Detection Accuracy** (proportion of true anomalies detected), **False Alarms per Hour (FA/h)** (reliability metric), and **Precision–Recall AUC** (PR-AUC) for robustness on imbalanced data.

5.2 Results and Analysis

GPS Trajectory Anomalies (LSTM Autoencoder): The LSTM autoencoder effectively captured temporal dependencies in vessel trajectories, as shown in Figure 1. The simulated vessel trajectory, together with the GPS anomalies identified by the LSTM autoencoder, is visualised. The underlying track (shown in light blue)

represents the vessel's nominal movement pattern, while the red markers highlight locations where the model flagged deviations from expected behaviour. These anomalies occur in spatial clusters, indicating periods where the vessel's motion diverged from its learned trajectory dynamics, consistent with spoofing, drift, or injected positional perturbations. Because the LSTM autoencoder reconstructs the expected temporal evolution of latitude–longitude sequences, abrupt or sustained inconsistencies result in elevated reconstruction errors that allow the model to isolate abnormal segments of the track.

The model achieved a detection accuracy exceeding 94.2% with a PR-AUC of 0.91, typically identifying anomalous behaviour within 5-8 minutes of its onset. This delay reflects the need for gradual deviation patterns to accumulate sufficient temporal discrepancy before being distinguished from legitimate maneuvering. The visualization clearly demonstrates how the system localizes and clusters GPS anomalies along the vessel's route, supporting real-time situational awareness for maritime monitoring.

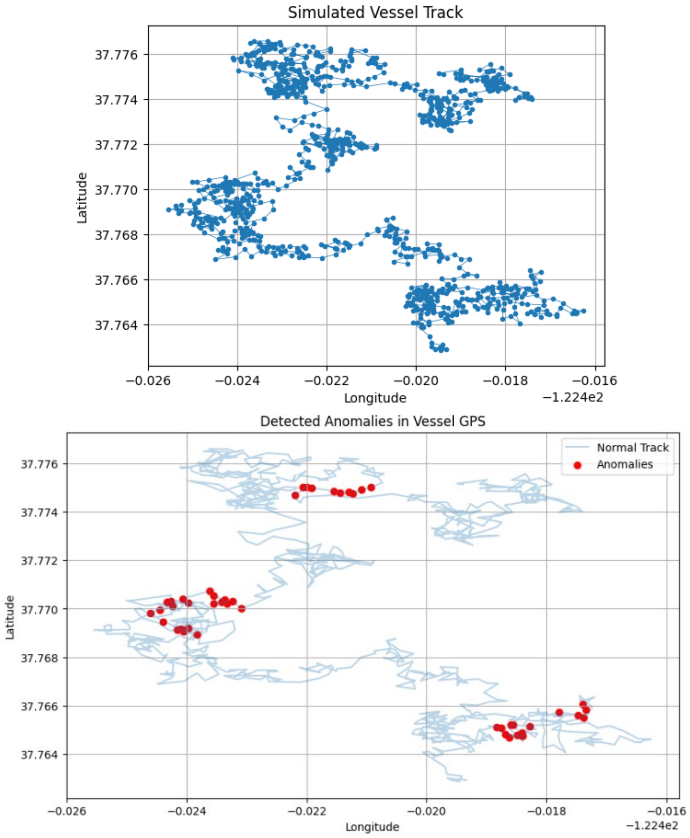


Figure 1: Detected GPS Trajectory Anomalies

Fuel level and engine temperature results, as shown in Figure 2, illustrate the behaviour of both fuel-level measurements and engine-temperature readings over time, with anomalous points highlighted in red. The fuel-level plot shows several abrupt drops to zero, indicating sensor tampering, data corruption, or sudden, unrealistic depletion events. Similarly, the engine-temperature plot reveals sharp downward spikes to zero and occasional extreme upward bursts, all of which deviate significantly from normal engine operating ranges.

The Isolation Forest model successfully identified each of these deviations as outliers, demonstrating its effectiveness in capturing both instantaneous spikes and sustained abnormal values in OT sensor streams. Detection accuracy reached 92.8%, with a very low false-alarm rate (FA/h < 0.12). Furthermore, the model exhibited a short detection latency (typically under 4 minutes) due to the sharp and non-temporal nature of these anomalies, which the Isolation Forest is well-suited to identify quickly. These results confirm the model's ability to reliably detect fuel and engine abnormal behaviours, as summarised in Table 2.

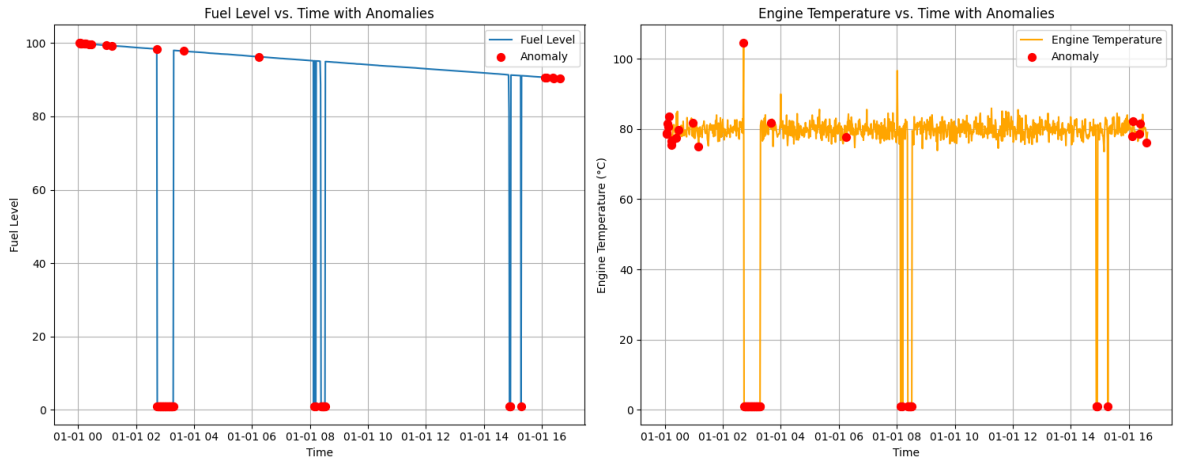


Figure 2: Detected Fuel and Engine Sensor Anomalies

The data, when visualised as a time-series plot, displays the temporal distribution of detected anomalies grouped into four distinct clusters based on their occurrence times, as shown in Figure 3. Each cluster represents a period in which multiple anomalies occurred in close succession, separated from other clusters by longer inactive intervals. Cluster 0 corresponds to the earliest anomaly window, while Cluster 1 captures a sequence of mid-range anomalies. Cluster 2 highlights a later group of anomalies, and Cluster 3 captures the final anomalies.

The clustering pattern illustrates how anomalies tend to emerge in grouped time segments rather than as isolated events, indicating periods of concentrated unusual system behaviour. This analysis supports the system's ability to not only detect individual anomalous readings but also characterise broader anomaly episodes, contributing to more contextual and situationally aware maritime monitoring.

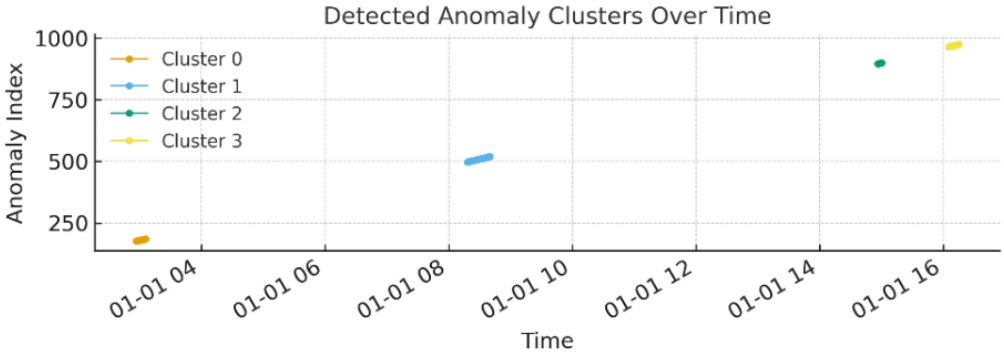


Figure 3: Detected Anomaly Clusters Over Time

5.3 Model Performance Comparison

Table 2: Performance summary of anomaly detection models across modalities.

Model	Data Type	Detection Accuracy	PR-AUC	FA/h	Detection Latency
LSTM Autoencoder	AIS / GPS (temporal)	94.2 %	0.91	0.14	5–8 min
Isolation Forest	Fuel & Temp Sensors	92.8 %	0.88	0.12	< 4 min
Isolation Forest	Compass Signals	90.3 %	0.85	0.16	6–10 min

5.4 Insights and Operational Implications

The results demonstrate that modality-specific anomaly detection substantially enhances precision and interpretability across the vessel's navigation and operational data streams. The LSTM autoencoder effectively identifies trajectory-based inconsistencies (Figure 1), capturing subtle spatial deviations that emerge only through temporal sequence modelling. In contrast, the Isolation Forest excels at detecting abrupt fuel-level and engine-temperature irregularities (Figure 2), flagging sudden drops, spikes, and implausible readings with minimal delay. The temporal grouping of anomalies (Figure 3) further reveals that abnormal events tend to occur

in concentrated clusters rather than as isolated incidents, providing valuable contextual insight for operational monitoring.

Detection latency varies significantly by data type: operational sensor anomalies are detected fastest (<4 minutes), whereas trajectory deviations require 5-10 minutes to accumulate sufficient evidence for reliable classification. These findings underscore the value of a hybrid, multi-model architecture in which each modality is paired with the model best suited to its statistical structure. By deploying complementary models within a modular system, VessiGuard maximises anomaly-detection coverage while maintaining low false-alarm rates. The resulting system offers low latency, interpretable outputs, and clear visual diagnostics suitable for real-time maritime monitoring, enabling operators to respond rapidly and prevent escalation during navigation or incident handling.

6. Limitations and Future Work

6.1 Dataset Constraints and Adversarial Robustness

A primary limitation of this study is its reliance on simulated and publicly available datasets. Although sufficient for proof-of-concept validation, these datasets do not fully capture the complexity and noise characteristics of real-world maritime environments. Furthermore, the current evaluation did not include adversarial robustness testing. Consequently, the system's resilience against intentionally crafted and adaptive cyberattacks, specifically designed to mislead machine learning models, remains to be validated. This limits the certainty of performance under sophisticated, targeted threat scenarios.

6.2 Operational Integration and Crew Impact

The current implementation was evaluated offline and not integrated into a real-time bridge system. Therefore, critical operational factors such as **integration challenges with existing navigation equipment**, onboard computational latency, and the **impact of false alarms on crew workload** were not empirically tested. Future deployment must also consider a **cost-benefit analysis** for operators, weighing the expense of retrofitting vessels with edge-computing hardware against the potential risk reduction. Addressing these barriers is essential to ensure the system supports, rather than distracts, vessel crews.

6.3 Future Directions

Future work will transform VessiGuard from a validated prototype into a deployable solution:

- Real-World Deployment and Sensing Expansion: Edge deployment on operational vessels to evaluate performance under bandwidth constraints and incorporate radar, weather, and vessel-to-infrastructure data for enhanced situational awareness.
- Adversarial Robustness Testing: Conducting rigorous stress-testing using adversarial samples to quantify and improve the model's resilience against adaptive evasion attacks.
- Human-Centred XAI Dashboard: Developing a regulation-compliant operator interface with Explainable AI (XAI). This will mitigate **alarm fatigue** by providing context-aware visualisations that help crews quickly distinguish between sensor faults and genuine cyber threats.
- System Integration and Advanced Learning: Implementing secure logging for immutable incident audit trails and exploring hybrid architectures (combining supervised and reinforcement learning) to improve adaptability across diverse vessel types.

7. Conclusion

This research presented VessiGuard, an intelligent anomaly detection system tailored to the evolving cybersecurity needs of the maritime sector. By combining Long Short-Term Memory (LSTM) models for sequential navigational **behaviour** and Isolation Forest algorithms for operational sensor anomalies, the system effectively identifies a wide range of threats, including GPS spoofing, sensor manipulation, and irregular vessel **behaviour**. The results validate the feasibility of modality-specific detection strategies, demonstrating strong detection performance and low false alarm rates. However, this work should be understood as a proof-of-concept, validated using simulation-based scenarios. Beyond its technical contributions, VessiGuard offers a foundational pathway toward adaptive maritime cybersecurity. While currently a prototype, its modular architecture and potential for future integration with existing vessel systems provide a blueprint for industry adoption. As the maritime domain continues its digital transformation, solutions like VessiGuard will be essential for safeguarding shipping infrastructure, ensuring vessels operate securely and resiliently in an increasingly complex threat landscape.

AI Declaration: Artificial intelligence and language support tools were used to prepare this manuscript. OpenAI's ChatGPT was utilised to assist with language refinement, structure enhancement, and technical phrasing, while Grammarly was employed for maintaining style consistency. The authors conducted all conceptualisation, system design, methodology, data analysis, and results interpretation. All AI-generated outputs were critically reviewed and verified for accuracy before inclusion.

Ethics and AI Declaration: This study did not involve human participants, sensitive personal data, or interventions requiring ethical approval. All data used were publicly available or synthetically generated within a controlled simulation; consequently, no institutional ethics clearance was required.

References

Amro, A., Oruc, A., Gkioulos, V. and Katsikas, S. (2022). 'Navigation data anomaly analysis and detection', *Information*, 13(3). Available at: https://www.mdpi.com/2078-2489/13/3/104

BIMCO, CLIA, ICS, INTERCARGO and INTERTANKO (2020). *The Guidelines on Cyber Security Onboard Ships (Version 5).* BIMCO, Bagsværd, Denmark.

Bobrovnikova, K., Lysenko, S., Hurman, I. and Kwiecien, A. (2022). 'Machine learning based techniques for cyberattack detection in the Internet of Things infrastructure', in *IntelITSIS*, pp. 411–420.

Bolbot, V., Kulkarni, K., Brunou, P., Banda, O.V. and Musharraf, M. (2022). 'Developments and research directions in maritime cybersecurity: A systematic literature review and bibliometric analysis', *International Journal of Critical Infrastructure Protection*, 39, 100571.

Brenner, B., Hollerer, S., Bhosale, P., Sauter, T., Kastner, W., Fabini, J. and Zseby, T. (2023). 'Better safe than sorry: Risk management based on a safety-augmented network intrusion detection system', *IEEE Open Journal of the Industrial Electronics Society*, 4, pp. 287–303.

Davari, N. and Aguiar, A.P. (2021). 'Real-time outlier detection applied to a Doppler velocity log sensor based on hybrid autoencoder and recurrent neural network', *IEEE Journal of Oceanic Engineering*, 46(4), pp. 1288–1301.

Golosova, J. and Romanovs, A. (2018). 'The advantages and disadvantages of the blockchain technology', in *IEEE 6th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, pp. 1–6.

Hochreiter, S. and Schmidhuber, J. (1997). 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735–1780.

Hu, Q., Han, W. and Zhang, H. (2021). 'Ship identity authentication security model based on blockchain', in *4th International Conference on Data Science and Information Technology*, pp. 135–142.

IMO (2017a). *Guidelines on maritime cyber risk management (MSC-FAL.1/Circ.3).* International Maritime Organization, London.

IMO (2017b). *Maritime cyber risk management in safety management systems (MSC.428(98)).* International Maritime Organization, London.

Kim, H. and Joe, I. (2024). 'Enhancing anomaly detection in maritime operational IoT time series data with synthetic outliers', *Electronics*, 13(19). Available at: https://www.mdpi.com/2079-9292/13/19/3912

Liu, F.T., Ting, K.M. and Zhou, Z.-H. (2008). 'Isolation forest', *IEEE International Conference on Data Mining*, pp. 413–422.

Maganaris, C., Protopapadakis, E. and Doulamis, N. (2024). 'Outlier detection in maritime environments using AIS data and deep recurrent architectures', *CoRR*, abs/2406.09966. Available at: https://doi.org/10.48550/arXiv.2406.09966

Nguyen, D., Simonin, M., Hajduch, G., Vadaine, R., Tedeschi, C. and Fablet, R. (2020). 'Detection of abnormal vessel behaviours from AIS data using GeoTrackNet: From the laboratory to the ocean', in *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, pp. 264–268.

Ribeiro, C.V., Paes, A. and de Oliveira, D. (2023). 'AIS-based maritime anomaly traffic detection: A review', *Expert Systems with Applications*, 231, p. 120561.

Riveiro, M., Pallotta, G. and Vespe, M. (2018). 'Maritime anomaly detection: A review', *WIREs Data Mining and Knowledge Discovery*, 8(5), e1266.

School of International and Public Affairs, Columbia (2022). *NotPetya: Lessons learned from the first global cyberattack on critical infrastructure.* Available at: https://www.sipa.columbia.edu/sites/default/files/2022-11/NotPetya%20Final.pdf

Talukder, M.A., Islam, M.M., Uddin, M.A., Hasan, K.F., Sharmin, S., Alyami, S.A. and Moni, M.A. (2024) 'Machine learning-based network intrusion detection for big and imbalanced data', *Journal of Big Data*, 11(1), p. 33.

Varma, R. (2024). *VessiGuard Synthetic Maritime Sensor and Navigation Dataset (Version 1.0).* Unpublished raw data generated by the author.

Wang, W., Bin, J., Zaji, A., Halldearn, R., Guillaume, F., Li, E. and Liu, Z. (2022). 'A multi-task learning-based framework for global maritime trajectory and destination prediction with AIS data', *Maritime Transport Research*, 3, p. 100072.

The Informed Fake News Advisor (IFNA): Toward Sociotechnical Solutions for Fake News Detection

Namosha Veerasamy and Danielle Badenhorst
CSIR, Pretoria, South Africa

nveerasamy@csir.co.za
dbadenhorst@csir.co.za

Abstract The detection of fake news is a multidimensional challenge that demands solutions extending beyond purely computational approaches. Although advances in natural language processing, machine learning, deep learning, and multimodal analysis have strengthened technical capabilities, misinformation continues to proliferate. Fake news thrives within environments shaped by complex social interactions, platform-specific advantages, and human judgement. Social factors (such as user profiles, sharing behaviours, engagement metrics, network structures, and crowd-sourced credibility signals) play a critical role in how misinformation is created, propagated, and perceived, yet these contextual nuances are often overlooked in algorithmic models. These social dimensions operate alongside technical elements, including linguistic cues, visual content, temporal dissemination patterns, and hybrid feature integration. Drawing on a review of the recent literature, this work synthesises sociotechnical elements to inform the development of an integrated approach to the detection of fake news. We introduce the SHAPE conceptual framework to guide the development of the Informed Fake News Advisor (IFNA). This conceptual framework will guide the creation of IFNA, which will consist of detection tools that combine technical precision with sensitivity to social context. By framing fake news detection as a sociotechnical problem, IFNA shifts the focus from isolated technical optimisation towards a holistic design philosophy, supporting the development of solutions that are both effective in detection and responsible in deployment within complex information ecosystems.

Keywords: Fake news, Fake news detection, Sociotechnical systems, Framework

1. Introduction

Misinformation can include a range of information reliability that moves from verified facts to distortion (e.g., exaggeration and selective use of evidence), or rumour (maybe true and maybe false), and finally to outright falsehoods (lies and fake news) (Hendricks and Vestergaard, 2018). Relative concepts related to fake news include satire, yellow journalism, junk news, pseudo-news, hoax news, propaganda news, advertorial, false information, fake information, misinformation, disinformation, mal-information, alternative fact, and post-truth (Wang, 2020).

Figure 1 shows the relationship between many terms relevant to fake news. Aïmeur et al. (2023) provide the following definitions:

- Misinformation is false information that is shared without the intention of misleading or causing harm,
- Disinformation is false information that is shared to intentionally mislead,
- Malinformation is genuine information that is shared with the intention of causing harm.

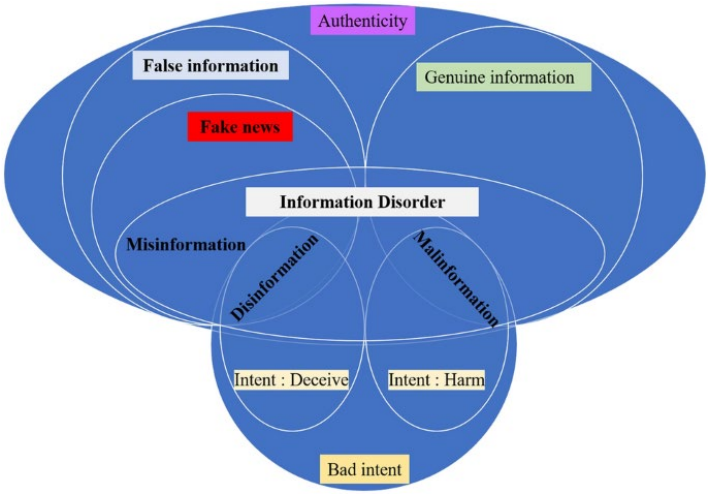


Figure 1: Relationship of Terms Related to Fake News (Aïmeur et al., 2023)

Stanford University defines fake news as news articles that are intentionally and verifiably false and could mislead readers (Allcott and Gentzkow, 2017).

Fake news entails the use of false stories and news that may mislead or lure readers for financial, political, or other purposes. Fake news can be harmful for many reasons (Shu et al., 2019):

- People are misled and accept false beliefs
- Changes the way people respond to true news
- Breaks the trustworthiness of the entire news ecosystem

Fake news has become a significant problem, as the intentional spread can lead to misleading information being mistaken for genuine news. Due to the widespread use of digital technologies, such as the Internet and social media, fake news can proliferate at a rapid rate. Fake news has become a significant problem, as the intentional spread of misinformation can lead to misleading information being mistaken for genuine news. Additionally, the use of digital technologies, such as the Internet and social media, can lead to the rapid proliferation of fake news.

Addictive news platforms lack editorial controls, allowing wide dissemination of misinformation via algorithmic news feeds, and without the necessary digital media literacy skills, people frequently fall victim to dubious claims they encounter in this context (llaire-Duquette et al., 2025). Fake news can mislead the public, create distorted perceptions, and negatively influence decision-making. Meanwhile, humans have been shown to be not proficient in differentiating between truth and falsehood when overloaded with deceptive information (Zhou et al., 2020). The sheer volume of information that users are overwhelmed with makes it difficult to detect fake news.

An additional significant challenge in these contemporary information ecosystems is the increasing use of social media to access news. This phenomenon is particularly acute in regions such as South Africa, where 73% of respondents report concern about their ability to discern true from false information online: a figure comparable to Nigeria and the United States, and substantially above the global average of 58% (SANEF, 2025). In the United States, 38% of adults regularly obtain news from Facebook and 35% from YouTube, underscoring the reliance on platforms that often lack robust editorial oversight (Pew Research Center, 2025). Meanwhile, in South Africa, 33% of users access news via TikTok and 42% via YouTube, among the highest rates on the continent (SANEF, 2025).

Fake news can also take many different forms, such as fake reviews, false political statements, and misleading advertisements. This affects the ways that people make decisions, can impact brand reputation and can also damage consumer trust. Fake news can sometimes be challenging to detect due to ambiguities in language and the deliberate crafting of content that mimics the tone and structure of genuine news. Detection algorithms can sometimes fail to detect these subtle nuances, including the use of satire, humour, and opinion writing. Such pieces do not adhere to journalistic practices, such as objective reporting. The manipulation of images and video clips can make fabricated stories appear more convincing. Clickbaiting techniques also draw the readers in with misleading and sensational headlines.

Contemporary detection systems face substantial constraints; technical challenges may impact the effectiveness of fake news detection systems. If algorithms are trained on datasets that are not large and diverse enough, bias can be introduced. The limited availability of high-quality training data for fake news detection is a big issue for training supervised learning models (Zhang and Ghorbani, 2020). This will impact the ability of these tools to detect fake news. Automated detection of fake news is simply not enough to flag content. Machine models can help identify patterns, but they can also overlook clever uses of language that aim to evade detection. Confirmation biases, echo chambers, and cognitive dissonance remain challenges, as fact-checking can still be rejected. New strategies for creating fake news will also emerge. Bots can be used to create fake news, which can then be disseminated. Online users may also not take accountability for content that is posted, shared and commented on. Inadvertently, they also contribute to the proliferation of the problem.

The acceptance and proliferation of fake news reflect complex interactions between social, psychological, and technical factors. Existing resources for protecting against misinformation demonstrate limited efficacy, depending on individuals' ability to discriminate, corroborate, and verify the authenticity of information. Technocentric approaches alone prove insufficient to address this multidimensional challenge; a more comprehensive framework that integrates human cognition, social dynamics, and technical capacity is necessary.

From a cybersecurity perspective, Caramancion et al. (2022) argue that disinformation, including fake news, can be framed as a cybersecurity threat as it exploits human cognitive vulnerabilities to compromise the integrity, availability, and trustworthiness of information, analogous to traditional cyberattacks on systems or networks. Recognising this as a cyber threat prompts the need for socio-technical defences addressing both human- and system-level vulnerabilities.

The contribution of this work is therefore not the introduction of a novel detection algorithm, but the development of a design-oriented framework that synthesises fragmented research across cognitive psychology, social network analysis, and computational fake news detection. By framing fake news detection as a sociotechnical design problem, this paper provides a theoretical foundation and research agenda for the development of more effective and responsible detection systems.

The rest of the paper is structured as follows: Section 2 reviews current approaches to fake news detection and proliferation, discussing cognitive theories, social-networks-based approaches, and technical/computational approaches. Section 3 proposes the SHAPE model, which helps guide the user in techniques for dealing with information processing. Section 4 provides a discussion, limitations, and future work. Section 5 concludes the paper.

2. Current Approaches to Fake News Detection and Proliferation

Fake news proliferation detection methods have been studied in various disciplines. Techniques range from cognitive theories and approaches (Section 2.1), social network-based approaches (Section 2.2), and technical or computational approaches (Section 2.3). Fake news detection methods can rely on a technocentric paradigm only, which has its limitations (Section 2.4).

2.1 Cognitive Theories and Approaches

Psychology plays a fundamental role in shaping how we process information based on the reinforcement of certain beliefs and our current emotional state. Cognitive biases, emotional responses, and motivational factors are deeply embedded in our psychological functioning. Forms of confirmation bias include (University of California Irvine, 2025)

- Familiarity Bias: The repetition of misinformation makes it feel more believable simply because people are exposed to it frequently, a phenomenon enhanced by social media's high-volume information flow.
- Availability Bias: People tend to believe information that is more easily recalled or accessible in memory, which often favours sensational or emotionally charged fake news stories.
- Confirmation bias: People may tend to strongly believe and share fake news when it supports their pre-existing beliefs

Biases are mental shortcuts that reduce the cognitive effort required to process information. However, it also makes us vulnerable to fake news. Biases, when combined with emotional responses, social pressures, and digital algorithms, can reinforce prior beliefs and predispose users to believe fake news due to the way content is presented. Our interpretation of fake news is influenced by our existing beliefs, emotional state, and social priorities, as our brains tend to opt for the fastest route to processing information without engaging in proper critical thinking and fact-checking processes.

When humans experience emotions such as fear, joy, anxiety, or anger, it can affect their ability to evaluate the credibility of information and share content without proper verification. Deeply rooted in our psyche is the need for gratification, autonomy, social acceptance, and altruism; Sharing content satisfies these needs, even if people are being deceived or sharing misleading information. This form of fast sharing for approval, inclusion, fulfilment, and entertainment has created a society where emotionally charged and sometimes biased or sensationalised content is spreading rapidly on various digital platforms.

Fake news can be used as a form of emotional manipulation to amplify feelings of anger, joy, anticipation, or excitement. The use of sensationalist or emotionally charged expressions/imagery can trigger emotional reactions that override rational and critical thinking. When users are in emotional states such as anger, happiness, excitement, or agitation, they may be more susceptible to believing fake news, as their judgement can be clouded by the intense emotional responses they are experiencing. A lack of education and scientific knowledge can influence some individuals to be more susceptible to fake news. Lower education and knowledge

levels are associated with increased engagement with fake news on social media, such as clicking links or sharing posts, thus helping to spread misinformation (Mihai-Ionuț and Irina, 2019).

Some of the cognitive factors that have influenced the spread of fake news include (Beauvais, 2022):

Confirmation bias, where individuals seek information that aligns with their existing beliefs,

Political partisanship affecting political beliefs,

Prior exposure and the illusory truth effect, whereby people believe a statement to be true as it is repeated frequently,

Type of reasoning: intuitive thinking vs. analytic thinking,

Lack of critical thinking,

Little time spent on news,

To be "in the know",

Low educational level,

Low science knowledge, and

Trust in "elite" messages.

Beauvais (2022) also mentions the following psychological factors that have led to the spread of fake news:

- Attraction to novelty and emotionally evocative content, often leading to impulsivity in sharing,
- High emotional states and low emotional intelligence can increase susceptibility,
- Anti-science attitudes also play a role.

Shane (2020) also discusses some of the key psychological reasons that influence our susceptibility to misinformation (and fake news).

- Cognitive miserliness: Humans prefer to use simpler and easier ways of solving problems that require the least mental effort. This is part of what makes our brains so efficient: we do not want to think very hard about everything. However, this also leads us not to put in enough effort into things that may require more effort, especially when we encounter news stories and fail to use critical thinking to first evaluate the content.
- Dual process theory: There are 2 basic ways of thinking. System 1 is an automatic process that requires little effort, and System 2 is an analytical process that requires more effort. Since we are cognitive misers, we tend to default to System 1 (easy form) thinking. Automatic processing is a simpler and faster approach; we tend to make easy judgements, but it can be inefficient, as we can miss crucial details.
- Heuristics are used as indicators for quick judgments instead of complex analysis. For example, relying on a social endorsement heuristic, such as a post on a social media site that appears to come from someone who can be trusted. However, the post could have been falsely created to impersonate the person and may not be a completely reliable indicator that the news story is true.
- Cognitive dissonance occurs when a person encounters information that contradicts beliefs, which may result in the rejection of credible information to alleviate the dissonance.
- Motivated reasoning: It is when people use reasoning skills to believe what they want to believe, instead of determining the truth. People's rational faculties can lead to misinformed beliefs, rather than lazy thinking.
- Pluralistic ignorance: a lack of understanding about what others in society think and believe. This can lead people to incorrectly believe that others share a particular political view when, in fact, it is held by a very small number of people. A related phenomenon is the false consensus effect, in which people overestimate the number of others who share their views.

Kahnemann (2011), in his renowned book "Thinking, Fast, and Slow", explains that the way to block errors that originate in System 1 is simple in principle: recognise the signs that you are in a cognitive minefield, slow down, and ask for reinforcement from System 2 (Kahnemann, 2011). Saltor et al. (2023) showed that reflective thinking styles and actively open-minded thinking predicted better fake news discrimination, with actively open-minded thinking being the strongest predictor. Saltor et al. (2023) explain that reflectiveness, that is, conscious and

deliberate thinking processes, can intervene in the fast default responses provided by intuitive thinking if prompted. The human brain is naturally inclined toward rapid, low-effort cognitive processing, which increases reliance on intuitive rather than analytical thinking. However, the key to preventing the propagation of fake news will be to slow down automatic thinking processes and apply more deliberate, critical thinking. This will be further discussed in the conceptual IFNA framework to help guide the user on principles for the prevention of fake news.

2.2 Social Network Approaches

Fake news is predominantly distributed using social media platforms or Online Social Networks (OSNs); this mode of news diffuses faster and more broadly than the truth (Vosoughi et al., 2018). Additionally, OSNs maximise engagement; this skews content visibility, favouring posts similar to previously engaged content while suppressing alternative perspectives. Critically, this approach has pointed to negative societal consequences, potentially reducing content diversity and amplifying divisive material (Bouchaud, 2024). The' unique characteristics of the platforms contribute to this issue: unlike traditional media, social media users do not choose their news sources, instead relying on proprietary algorithms that provide targeted, often non-transparent information (Shu et al., 2017).

Within these OSNs, connections on social media form rich, interconnected graph structures. These graphs can represent users or content as nodes, connected by edges (links) that represent various relationships, such as content similarity, reposts, friendships, etc. (Lakzaei et al., 2024). Utilising these graphs, Social Network Analysis (SNA) approaches can utilise the structural dimensions of these sharing networks, behaviour patterns, and interactions, to provide additional context to solely content-based methods (as discussed in Section 3.3). Furthermore, SNA can be applied at varying network levels (node-level, ego-level, triad-level, community, or whole network), allowing rich information and analysis opportunities (Zhou and Zafarani, 2019).

2.2.1 Graph-based detection methods

Graph-based detection uses relationship modelling between news articles, users, content, and external sources as graphs. This includes techniques such as Propagation-Based Methods (Section 2.2.1.1), Knowledge-Driven Methods (Section 2.2.1.2), Social Context and Heterogeneous Graphs (Section 2.2.1.3) and Multi-Modal and Evidence-Aware Graphs (Section 2.2.1.4).

- Propagation-Based Methods

Broadly, these deep-learning methods utilise Graph Neural Networks (GNNs) to consider how fake and real news propagate differently. Fake news spreads through social networks differently in comparison to other forms of content. Using graph structures, such as network centrality (the degree to which a node is central to a network), it is possible to identify indicators such as: (a) rapid dissemination, (b) clustering, and (c) influential nodes that are associated with fake news (Sivasankari and Vadivu, 2022).

GNNs have proven particularly effective in capturing propagation patterns, frequently outperforming text-only baselines, especially in early detection scenarios and in multilingual contexts where language-agnostic approaches are required (Han et al., 2021; Harby and Zulkernine, 2023). The temporal and dynamic nature of information diffusion has motivated the development of temporal-delay GCNs and dynamic graph snapshots, which capture the evolving topology of network structures over time (Song et al., 2022, 2021). These temporal extensions address limitations of static graph approaches: these systems can model how information flows evolve as fake and real news traverse networks at different velocities and scales. Various GNNs have been developed, including Graph Convolution Neural Network (BiGCN-A, BiGCN-B), Graph Attention Neural Network (BiGAT), GraphSage (BiSAGE), Graph Convolution with ARMA filters (BiARMA), and Simplified Graph Convolution Neural Network (BiSGCN) (Harby and Zulkernine, 2023), with varying levels of accuracy.

- Knowledge-driven Methods

Knowledge-driven Methods construct knowledge graphs by extracting entities and relationships from news content and cross-referencing them against external databases such as Wikidata. This method frames fake news detection as a fact-checking problem, wherein claims extracted from news articles are compared against verified factual assertions using graph matching or embedding-based techniques (KEHGNN-FD, DEAP-FAKED, MiLk-FD) (Mayank et al., 2022; Phan et al., 2023; Xie et al., 2024). Heterogeneous knowledge graphs (integrating diverse entities, such as news articles, semantic entities, topics, and/or factual claims) can provide rich representations in comparison to homogeneous graphs. Heterogeneity can introduce nuanced reasoning about the relationships

between claims and external knowledge and has resulted in improved detection precision in the literature (Xie et al., 2024).

- Social Context and Heterogeneous Graphs

Beyond propagation patterns and factual verification, heterogeneous graph methods model complex interactions among users, news content, comments, and topics as integrated graph structures. Representative approaches include DAGA-NN (Domain-Adversarial Graph Attention Neural Networks) and Graph Attention Networks (GATs) (Chang et al., 2024; Yuan et al., 2021). These approaches capture multiple dimensions of credibility assessment simultaneously: user reputation and historical reliability, community structure and clustering patterns, and cross-domain relationships that may indicate coordinated inauthentic behaviour (Phan et al., 2023). Graph Attention mechanisms and domain-adversarial networks enable these methods to weight the significance of different node and edge types, allowing the model to learn which social signals are most indicative of fake news (Yuan et al., 2021).

- Multi-Modal Graphs

Contemporary detection methodologies integrate multiple modalities: text, images, and external evidence into unified graph structures. Multi-modal graph approaches, such as TEMGNNs (Temporal Enhanced Multi-modal Graph Neural Networks), MEFaND, and SAFE (Social Attention Fusion Engine), use GNNs to integrate features across modalities and explicitly model semantic relationships between textual claims and visual content (Chang et al., 2024; Sormeily et al., 2024; Yuan et al., 2021). These methods address the credibility that manipulated images and videos lend to false narratives: these systems can better detect incongruities between claims and corroborating evidence that might be missed by unimodal approaches.

2.3 Technical Approaches

Fake news detection via computational means has garnered much attention in the literature. Considering the influx of information disseminated online, detecting fake news at scale computationally could address the limitations posed by manual fact-checking: a practice that is seemingly becoming impossible given the volume of digital content (Shah et al., 2024). Computational approaches that use machine learning and natural language processing (NLP) (Section 3.3.1), deep learning (Section 3.3.2), transformer models (Section 3.3.3), or ensemble systems (Section 3.3.4) to detect fake news with promising accuracy.

2.3.1 Machine learning and natural language processing

Machine learning techniques aim to automatically detect and classify fake news by analysing textual patterns and linguistic features. Traditional machine learning models such as logistic regression, decision trees, support vector machines (SVM), random forests, and Naive Bayes are commonly used for the detection of fake news. These models typically rely on features extracted from news content using NLP techniques such as tokenisation, lemmatisation, part-of-speech tagging, and term frequency-inverse document frequency (TF-IDF) (Prachi et al., 2022).

2.3.2 Transformer-based models

Transformer-based architectures have achieved state-of-the-art performance in fake news detection tasks, largely displacing earlier deep learning approaches. The foundational Bidirectional Encoder Representations from Transformers (BERT) model and its variants, including Robustly Optimised BERT Pre-training Approach (RoBERTa) and XLNet, have the capacity to capture semantic and contextual information via bidirectional attention mechanisms. These models benefit substantially from transfer learning, whereby pre-trained representations developed on large text corpora are fine-tuned on fake news detection tasks, reducing the requirement for task-specific training data (Aljawarneh and Swedat, 2024; Kaliyar et al., 2021; Raza and Ding, 2022).

- Large Language Models (LLMs)

LLMs have shown promising results in fake news detection through three primary mechanisms: (i) fake text classification, (ii) fact-checking and verification, and (iii) contextual analysis (Papageorgiou et al., 2024). Of particular interest is the ability to decompose claims, retrieve relevant evidence, and integrate sources, to minimise extensive human annotation. Techniques such as stepwise prompting and weak supervision have been shown to enhance accuracy and reduce hallucinations, previously constraining these methods (Kuntur et al., 2025). However, limitations are still present, as various fine-tuned BERTs still outperform LLMs in detection

accuracy, though LLMs are still used as "advisors" to provide multi-perspective instructive rationales (Hu et al., 2024).

2.3.3 Deep learning

Deep learning architectures have demonstrated substantially superior performance compared to traditional machine learning methods, particularly when applied to complex and large text corpora. Convolutional Neural Networks (CNNs) excel at capturing local patterns and n-gram features in textual data through hierarchical feature extraction. By applying convolutional filters to text sequences, CNNs identify frequently occurring lexical patterns characteristic of fake news (Kumar et al., 2020). Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks address a fundamental limitation of CNNs by explicitly modelling sequential dependencies and contextual information within news articles. Hybrid models that combine CNN and RNN, such as CNN-LSTM, RNN-CNN, and Bidirectional Gated Recurrent Unit-Bidirectional LSTM architectures (Bi-GRU-Bi-LSTM), thus incorporate both local and sequential information, often achieving substantially higher accuracy than unimodal approaches (Kumar et al., 2020; Nasir et al., 2021).

2.3.4 Ensemble systems

Ensemble learning methodologies combine predictions from multiple models to achieve improved robustness and accuracy, mitigating the individual weaknesses inherent in single-model approaches. Multimodal Ensembles represent a particularly sophisticated development, combining textual, visual, and social context features through integrated transformer-based and deep learning models (Ahmad et al., 2020; Ali et al., 2022; Hakak et al., 2021).

2.4 Current Limitations of Technocentric Approaches

Multiple elements limit the sole reliance on purely algorithmic or technocentric methods for the detection of fake news. The changing nature of misinformation tactics creates an adversarial challenge that static models cannot resolve, necessitating continuous retraining and deployment cycles to maintain performance as bad actors adapt to detection methods (Vishnupriya et al., 2024).

Furthermore, the lack of high-quality, diverse and community-driven datasets is a major barrier. Further compounding this is the dependence of the supervised learning technique on labelled data, which restricts scalability and adaptability. In low-resource settings or when considering data sets in different languages, this poses significant challenges (Klein and Wueller, 2017). Fundamental barriers can still constrain widespread applicability: this includes data bias in datasets, small or imbalanced datasets leading to under- or over-fitting concerns and limited empirically validated use-cases (Hamed et al., 2023).

3. SHAPE: The Conceptual Framework for IFNA

A more hybrid model is needed that combines technical methods of source verification and fact-checking with human oversight. AI tools can be used to identify potential fake news, but humans themselves will have influence and judgement. This section provides a more detailed discussion of combined cognitive, social network-based, and technical approaches for detecting fake news.

Previous work includes the IFAKT Framework, which proposes a fake news detection framework (Veerasamy et al., 2024). The recommended guidelines in the IFAKT framework include considering the following issues: cultivating awareness around the source by checking credibility and investigating unknown sources, and deconstructing information by analysing headlines, verifying facts, and identifying logical fallacies. Increase media literacy through cross-referencing and fact-checking, and practice critical thinking skills to minimise confirmation bias and foster lateral thinking. Lastly, building healthy habits through open discussion and being cautious, while remaining continually aware.

The BBC has also adopted the SIFT method to spot misinformation (Ruggeri, 2024). SIFT stands for

- STOP: do not automatically share the post, comment
- Investigate the source: Find out where the post came from. Conduct a web search, as the platform search results can be misleading and may appear to be from a reputable website. Investigate whether the source is a reputable media outlet. For sources involving individuals, the user must verify whether the individual is an expert in the subject. Determine the purpose of content stemming from businesses. Check if there are underlying objectives, such as sales, advertising, substantiating a donor, or political endorsement.

- Find Better Coverage: If, after checking the source, there are still some doubts, it may be necessary to do further analysis by using fact-checking services.
- Trace back to its original context: Even if a claim has been reported on a credit media outlet, it may not necessarily be original or may be taken out of context. Many people may make claims on social media, and posts can go viral. It is important to determine the original source of the claim.

To synthesise the various cognitive, social network, and technical elements of misinformation detection, a conceptual framework called SHAPE is proposed. The framework will lead to the future development of an Information Fake News Advisor (IFNA), which will consist of technical tools to assist users in identifying fake news.

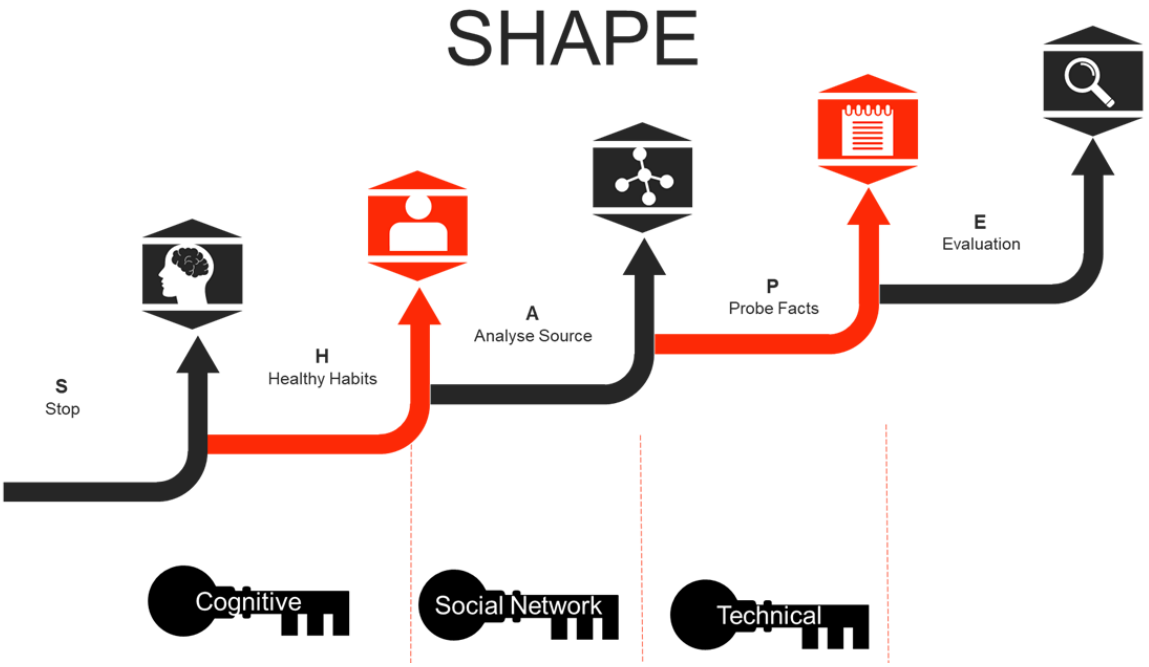


Figure 2: Conceptual Framework for IFNA

SHAPE (see Figure 2) is defined as:

S: Stop. Stop automatic processing but instead apply more lateral thinking and cognitive effort. Avoid the mental shortcuts of bias (familiarity, availability and confirmation bias). Users need to apply more analytical and critical thinking by using various socio-technical approaches. The aim is to think slower and not faster to actually analyse the content with various other cognitive processes, as well as social networks and technical techniques and tools.

H: Healthy Habits. Develop healthier habits to assess content and emotional responses. It is vital that users take a step back before sharing or reacting to determine whether the information is just trying to trigger a strong emotional response. As users utilise more socio-technical methods and tools to assess content, over time more healthy habits for information sharing can be developed. Users need to keep updated on media literacy and cybersecurity skills to learn about the spreading of misinformation and the latest cyber threats that may pose as real information.

A: Analyse source: Various platforms can be used to share and transfer information. These can include mainstream media, such as television, radio, and print, as well as digital social media platforms like Instagram, TikTok, Facebook, and Twitter/X. Users should verify the source of the content – is it a reputable site or author that can be trusted for fact-based information? Established news outlets, academic institutions and government agencies can be more trustworthy than random posts, headlines, adverts, reviews or claims. Cross-checking against other independent sources can also help verify the accuracy of stories.

Social-network detection approaches analyse sharing and behaviour patterns to identify certain indicators. Studying social-network detection approaches can help provide users with additional context, such as identifying influential nodes that spread the content, as well as insights into clustering and rapid dissemination. By

employing social-network detection techniques such as propagation-pattern analysis, network-based user credibility assessment, hybrid network content integration, and platform considerations, users can gain insight into how certain posts are favoured and others are suppressed. This can help users better assess the nature of the content.

P: Probe facts: The use of fact-checking websites can assist in verifying stories and debunking false claims. Perform a deeper headline analysis. Consider whether sensationalistic tactics have been applied with excessive emotion and punctuation. Users should reflect on whether the content has logical fallacies-is the story logical or feasible in the relevant context. Technical and computational methods can be used for the detection of fake news through the analysis of textual patterns and linguistic analysis.

E: Evaluate. Fundamentally, humans have at their disposal a set of socio-technical approaches as well as their human intuition to evaluate and make a judgement call. The key aspects will be to ask logical questions, examine the source, and determine whether the information is exaggerated and unlikely. Users can utilise technical techniques to assess the content, consult fact-checking platforms, and leverage tools and apps to help flag potential fake news. Overall, users need to develop stronger practices to cultivate a more discerning eye when faced with the flood of online content. There is no silver bullet for fake news detection, but with practice, discerning practices, and more critical thinking, users can develop better responses to the inundation of fake news.

4. Discussion

This study proposes sociotechnical integration as a necessity for scalable fake news detection. Technocentric approaches, while advancing detection accuracy, inevitably encounter limitations rooted in their methodologies: machine learning systems cannot anticipate novel adversarial strategies, graph-based methods depend on rich network data that are often unavailable on emerging platforms, and transformer models remain insensitive to the emotional and social contexts that motivate sharing behaviour. In contrast, psychological or cognitive approaches lack the scalability and precision necessary to detect false content at the current pace of dissemination.

The proposed SHAPE framework thus argues for the integration of technical tools and human judgement as complementary components within a sociotechnical system. When considering limitations, the framework remains conceptual and requires empirical validation through user studies and iterative refinement. Moreover, this work adopts a largely individualistic framing, emphasising user-level cognition; deeper systemic interventions addressing platform architectural choices, which are deliberately optimised for engagement rather than veracity, remain underexplored. Future work could involve empirical validation within the broader IFNA toolkit through controlled user studies. As a conceptual framework, SHAPE requires empirical validation. Future work will involve qualitative user studies and quantitative assessment of IFNA-assisted decision-making, measuring improvements in fake news discrimination, confidence, and response behaviour relative to unaided judgement.

5. Conclusion

Fake news detection remains a multidimensional challenge. It is not an issue that can be addressed solely by technical or social approaches. This paper argues that current algorithmic methods, whilst increasingly sophisticated, cannot yet adequately address misinformation in isolation from human cognition. The SHAPE framework bridges this gap by integrating cognitive theories, social network analysis, and computational detection into a coherent sociotechnical approach, guiding users to modulate their information processing through deliberative thinking, awareness of emotional manipulation, source evaluation, evidence verification, and critical judgement. The empirical development and iterative refinement of IFNA will require collaboration between disciplines and communities. Ultimately, addressing fake news effectively demands not a single solution, but a collaboration of disciplines and techniques.

Ethics Declaration: No ethical approval was required for this literature study.

AI Usage Declaration: Generative artificial intelligence (AI), specifically large language models (LLMs), was used in a limited capacity to help articulate some of the authors' ideas, concepts and explanations. No tools were used to write original content or generate *de novo* text. AI tools used to aid the writing process include ChatGPT and Perplexity.

References

Ahmad, I., Yousaf, M., Yousaf, S., Ahmad, M.O., 2020. Fake News Detection Using Machine Learning Ensemble Methods. Complexity 2020.

Aïmeur, E., Amri, S., Brassard, G., 2023. Fake news, disinformation and misinformation in social media: a review. Soc. Netw. Anal. Min. 13.

Ali, A.M., Ghaleb, F.A., Al-Rimy, B.A.S., Alsolami, F.J., Khan, A.I., 2022. Deep Ensemble Fake News Detection Model Using Sequential Deep Learning Technique. Sensors 22.

Aljawarneh, S.A., Swedat, S.A., 2024. Fake News Detection Using Enhanced BERT. IEEE Trans. Comput. Soc. Syst. 11.

Allcott, H., Gentzkow, M., 2017. Social media and fake news in the 2016 election. Journal of Economic Perspectives.

Beauvais, C., 2022. Fake news: Why do we believe it? Joint Bone Spine.

Bouchaud, P., 2024. Skewed perspectives: examining the influence of engagement maximization on content diversity in social media feeds. J. Comput. Soc. Sci. 7.

Caramancion, K.M., Li, Y., Dubois, E., Jung, E.S., 2022. The Missing Case of Disinformation from the Cybersecurity Risk Continuum: A Comparative Assessment of Disinformation with Other Cyber Threats. Data (Basel). 7.

Chang, Q., Li, X., Duan, Z., 2024. Graph global attention network with memory: A deep learning approach for fake news detection. Neural Networks 172.

Hakak, S., Alazab, M., Khan, S., Gadekallu, T.R., Maddikunta, P.K.R., Khan, W.Z., 2021. An ensemble machine learning approach through effective feature extraction to classify fake news. Future Generation Computer Systems 117.

Hamed, S.K., Ab Aziz, M.J., Yaakub, M.R., 2023. A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. Heliyon.

Han, Y., Karunasekera, S., Leckie, C., 2021. Graph Neural Networks with Continual Learning for Fake News Detection from Social Media. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12892 LNCS.

Harby, A.A., Zulkernine, F., 2023. A Comparative Analysis of Graph Neural Networks for Fake News Detection. In: Proceedings - International Computer Software and Applications Conference.

Hendricks, V.F., Vestergaard, M., 2018. Reality lost: Markets of attention, misinformation and manipulation, Reality Lost: Markets of Attention, Misinformation and Manipulation.

Hu, B., Sheng, Q., Cao, J., Shi, Y., Li, Y., Wang, D., Qi, P., 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In: Proceedings of the AAAI Conference on Artificial Intelligence.

Kahnemann, Daniel., 2011. Thinking, fast and slow. Farrar, Straus and Giroux.

Kaliyar, R.K., Goswami, A., Narang, P., 2021. FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. Multimed. Tools Appl. 80.

Klein, D.O., Wueller, J.R., 2017. Fake news: A Legal Perspective. Journal of Internet Law 20.

Kumar, S., Asthana, R., Upadhyay, S., Upreti, N., Akbar, M., 2020. Fake news detection using deep learning models: A novel approach. Transactions on Emerging Telecommunications Technologies 31.

Kuntur, S., Wróblewska, A., Paprzycki, M., Ganzha, M., 2025. Under the Influence: A Survey of Large Language Models in Fake News Detection. IEEE Transactions on Artificial Intelligence 6.

Lakzaei, B., Haghir Chehreghani, M., Bagheri, A., 2024. Disinformation detection using graph neural networks: a survey. Artif. Intell. Rev. 57.

llaire-Duquette, G., Hasni, A., Drouin, J.N., Groleau, A., Mahhou, A., Legault, A., Khayat, A., Carignan, M.E., Ayotte-Beaudet, J.P., 2025. Primary school pupils' ability to detect fake science news following a news media literacy intervention: Exploration of their success rate, evaluation strategies, self-efficacy beliefs, and views of science news. Journal of Digital Educational Technology 5, 2506.

Mayank, M., Sharma, S., Sharma, R., 2022. DEAP-FAKED: Knowledge Graph based Approach for Fake News Detection. In: Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2022.

Mihai-Ionuț, P., Irina, E.N.E., 2019. Influence of the educational level on the spreading of Fake News regarding the energy field in the online environment. In: Proceedings of the International Conference on Business Excellence. Sciendo, pp. 1108–1117.

Nasir, J.A., Khan, O.S., Varlamis, I., 2021. Fake news detection: A hybrid CNN-RNN based deep learning approach. International Journal of Information Management Data Insights 1.

Papageorgiou, E., Chronis, C., Varlamis, I., Himeur, Y., 2024. A Survey on the Use of Large Language Models (LLMs) in Fake News. Future Internet.

Pew Research Center, 2025. Social Media and News Fact Sheet [WWW Document]. URL https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/ (accessed 11.6.25).

Phan, H.T., Nguyen, N.T., Hwang, D., 2023. Fake news detection: A survey of graph neural network methods. Appl. Soft Comput.

Prachi, N.N., Habibullah, M., Rafi, M.E.H., Alam, E., Khan, R., 2022. Detection of Fake News Using Machine Learning and Natural Language Processing Algorithms. Journal of Advances in Information Technology 13.

Raza, S., Ding, C., 2022. Fake news detection based on news content and social contexts: a transformer-based approach. Int. J. Data Sci. Anal. 13.

Ruggeri, A., 2024. The "Sift" strategy: A four-step method for spotting misinformation [WWW Document]. BBC. URL https://www.bbc.com/future/article/20240509-the-sift-strategy-a-four-step-method-for-spotting-misinformation (accessed 10.28.25).

Saltor, J., Barberia, I., Rodríguez-Ferreiro, J., 2023. Thinking disposition, thinking style, and susceptibility to causal illusion predict fake news discriminability. Appl. Cogn. Psychol. 37.

SANEF, 2025. South Africa's Digital News Landscape in 2025 [WWW Document]. South African National Editors' Forum. URL https://sanef.org.za/south-africas-digital-news-landscape-in-2025/ (accessed 11.6.25).

Shah, N., Mitalia, N., Dhanare, R., 2024. Enhancing the Precision and Efficiency of Fake News Prediction: A Comprehensive Review. In: 2024 3rd International Conference on Sentiment Analysis and Deep Learning (ICSADL). pp. 335–341.

Shane, Tommy., 2020. The psychology of misinformation: Why we're vulnerable [WWW Document]. First Draft News.

Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H., 2017. Fake News Detection on Social Media: A Data Mining Perspective. ACM SIGKDD Explorations Newsletter 19, 22–36.

Shu, K., Wang, S., Liu, H., 2019. Beyond news contents: The role of social context for fake news detection. In: WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining.

Sivasankari, S., Vadivu, G., 2022. Tracing the fake news propagation path using social network analysis. Soft comput. 26.

Song, C., Shu, K., Wu, B., 2021. Temporally evolving graph neural network for fake news detection. Inf. Process. Manag. 58.

Song, C., Teng, Y., Zhu, Y., Wei, S., Wu, B., 2022. Dynamic graph neural network for fake news detection. Neurocomputing 505.

Sormeily, A., Dadkhah, S., Zhang, X., Ghorbani, A.A., 2024. MEFaND: A Multimodel Framework for Early Fake News Detection. IEEE Trans. Comput. Soc. Syst. 11.

University of California Irvine, 2025. Misinformation - Get the Facts [WWW Document]. https://guides.lib.uci.edu/Misinfo/bias.

Veerasamy, Namosha., Khan, Zubeida., Badenhorst, D., 2024. Examining Misinformation and Deep Fakes . In: Proceedings of the 25th European Conference on Knowledge Management. Academic Conferences International Limited, pp. 870–877.

Vishnupriya, G., Jeriel, A., Anish, R., Ajay, G., Giftson, A.J., 2024. Combating Fake News in the Digital Age: A Review of AI-Based Approaches. In: 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). IEEE, Pune, India.

Vosoughi, S., Roy, D., Aral, S., 2018. The spread of true and false news online. Science (1979). 359.

Wang, C.C., 2020. Fake news and related concepts: Definitions and recent research development. Contemporary Management Research 16.

Xie, B., Ma, X., Wu, J., Yang, J., Fan, H., 2024. Knowledge Graph Enhanced Heterogeneous Graph Neural Network for Fake News Detection. IEEE Transactions on Consumer Electronics 70.

Yuan, H., Zheng, J., Ye, Q., Qian, Y., Zhang, Y., 2021. Improving fake news detection with domain-adversarial and graph-attention neural network. Decis. Support Syst. 151.

Zhang, X., Ghorbani, A.A., 2020. An overview of online fake news: Characterization, detection, and discussion. Inf. Process. Manag. 57.

Zhou, X., Jain, A., Phoha, V. V., Zafarani, R., 2020. Fake News Early Detection. Digital Threats: Research and Practice 1.

Zhou, X., Zafarani, R., 2019. Network-based Fake News Detection. ACM SIGKDD Explorations Newsletter 21.

Rethinking the Human–Technical Split in Cybersecurity

Jukka Vuorinen
University of Jyväskylä, Finland

jukka.a.vuorinen@jyu.fi

Abstract: This paper re-examines one of the most enduring assumptions in cybersecurity and information systems: the split between the human and the technical. For decades, research and professional practice have portrayed the human as an unpredictable source of error—"the weakest link"—in contrast to the supposedly rational and controllable domain of technology. While this separation appears practical, it stems from a deeper lineage of Western thought that positions humans and nonhumans as fundamentally separate spheres. Drawing on thinkers such as Michel Foucault and Bruno Latour, this paper traces how this conceptual division has become embedded in security discourse, from early information systems design to contemporary frameworks such as ISO/IEC 27001 and the NIST Cybersecurity Framework. These standards institutionalize the split through parallel categories for "technical controls" and "human factors," shaping how security responsibilities are assigned and how failures are understood. The paper then explores what happens when this separation is challenged. Using examples such as intrusion detection systems and Trojan attacks, it shows that social and technical elements are inseparably mixed: anomalies, mimicry, and deception all rely on both code and conduct. Security decisions—from asset valuation to risk analysis—likewise emerge from socio-technical negotiations between what is desired and what is possible. To move beyond the limitations of this dichotomy, the paper introduces two frameworks that enact symmetry between human and technological agency. Conceptually, Actor–Network Theory treats both humans and artefacts as actors whose agency lies in their effects on others. Practically, Zero Trust security architectures operationalize the same symmetry by applying continuous verification equally to users and devices. Taken together, these perspectives suggest that cybersecurity should not be understood as two interacting domains but as a blended field of heterogeneous actors whose relations continually produce security. Recognizing this mixture does not dissolve the technical or the human but allows researchers and practitioners to see more clearly how each side folds into the other, reshaping what security can mean in practice.

Keywords: Cybersecurity, Human–Technical split, Sociomateriality, Actor–Network theory, Zero trust, Security frameworks

1. Introduction

Cybersecurity and Information Systems have long framed an apparent human–technical split. Typically, the story is that the human can shatter otherwise sturdy technical security (Baskerville, 1993; Loch, Carr & Warkentin, 1992). Later, the field became divided into the technical and the managerial, drawing in organizational and human aspects, with compliance emerging as a central concern (Siponen, Baskerville & Heikka, 2006; Bulgurcu, Cavusoglu & Benbasat, 2010). The division is often repeated in shorthand form: the human is the weakest link in security (Adams & Sasse, 1999; Nohlberg, 2009). This slogan performs a function: it re-establishes the split by suggesting that even if we had the strongest firewalls and most advanced technical solutions, these could be undone in an instant by a careless user.

This paper shows that the human–technical split in cybersecurity has, in fact, a strong historical context, yet the split can be counterproductive in practice, limiting and even disorienting. In terms of history, the split is not just a quirk of the field but part of a deeper lineage of Western thought. From Michel Foucault's (1970) account of the scientific order to Bruno Latour's (1993) views on modernity, there has long been a tendency in Western philosophy to separate human and nonhuman domains. The importance of paying attention to this split lies in the way it shapes our thinking, influencing what we perceive—and fail to perceive—about our environment and context. The same habit of thought informs cybersecurity: frameworks, standards, and professional practices reproduce the division, making it appear natural and necessary. Yet if we look more closely, the picture becomes more complicated in terms of how the technological and the human elements are set apart in the first place (Orlikowski, 2007). The sociomaterial turn is directly concerned with this very split (Orlikowski & Scott, 2008). If we examine the context of "the human is the weakest link," we can see that technical measures are saturated with social assumptions, while human practices are inseparable from the material infrastructures that enable them.

In this conceptual and theoretical essay, with illustrative examples, I examine three things. First, where the split comes from and how it has been inherited into the cybersecurity field, shaping how problems and responsibilities are framed in terms of technology and humans. Second, how the split is reproduced in practice—not only in popular sayings, but also in frameworks and standards that formally establish separate categories for technical and human factors (e.g. NIST, 2024; ISO/IEC 27001, 2022), even though at the level of actual practice the two are inseparably mixed. And third, how the boundaries can be approached differently.

On the operational side, Zero Trust architecture offers a case where humans and devices are treated in the same manner—each interaction checked continuously, no prior trust assumed or inherited. On the research side, Actor–Network Theory (ANT) (e.g. Latour, 2005) provides an analytic lens that refuses to recognize an ontological human–technical split, instead treating both as actors whose agency lies in their effects on others.

I do not argue that we should abolish categories like technical devices and human behavior, or deny that different types of controls (device configurations, architectures, or awareness training and compliance mechanisms) may be needed. Rather, I want to draw attention to how the split is constructed, why it has become so powerful, and how thinking differently might open new avenues for both researchers and practitioners. In other words, there is nothing inherently wrong with the split in many cases, but it can become a blinding factor if we imagine that technical solutions contain no human element or that human actors carry no technical dimensions. These are always blended. Zero Trust and ANT, taken together, provide ways to critically re-examine the division and to see cybersecurity not as two separate worlds, but as a blended field where humans and technologies constantly act on each other.

In this paper, the human–technical split refers not to any lack of integration between people and technologies in cybersecurity practice, but to an analytical distinction that treats human actors and technical artefacts as fundamentally different kinds of entities (see Orlikowski & Scott, 2008; Latour, 2005). This distinction is reproduced in cybersecurity discourse, standards, and frameworks that separate "human factors" from "technical controls" Accordingly, "social" denotes shared norms and institutionalized practices (not merely individual behavior or training), while the material/technical dimension denotes infrastructures, configurations, and algorithmic processes that enable and constrain action. Sociomateriality emphasizes that these dimensions are mutually constitutive in organizational practice.

2. Background – The Split Between Human and Technology

Saying that "the human is the weakest link" is frequently repeated in cybersecurity discourse (Adams & Sasse, 1999; Nohlberg, 2009). Whether this statement is accurate or not, it signals a widely shared view of a sharp split: on one side is the technical system while on the other is the human (user) side (e.g., Baskerville, 1993; Loch, Carr & Warkentin, 1992). These sides are seen as very different – isolated even: the former can be ruled by algorithms and is logical and predictable. Technology is programmable, but code is not impeccable. Mistakes do occur, but the code can be rewritten, corrected, and updated. Exact fixes are possible. Sometimes, however, updates in complex systems can go wrong—as in July 2024, when a CrowdStrike Falcon content update (Channel File 291) triggered a worldwide outage of Windows systems, requiring manual on-site remediation (CrowdStrike, 2024a; 2024b; CISA, 2024). Yet, systems are ultimately re-programmable and in that sense under control.

However, this split—the separation of human from others—is not unique to cybersecurity or information systems. It has a long history in how humans have understood the world and themselves, both in science and in everyday thought. Michel Foucault (1970), in The Order of Things, showed how scientific thinking developed around a special role for the human as the center of knowledge. Bruno Latour (1993), in contrast, argued that modernity itself rests on the repeated attempt to separate humans from nature, a distinction that is never fully achieved. These works reveal that separating the human from the nonhuman has been a recurring move in Western philosophy and science. Information Systems, despite often claiming to bridge the technical and the human, inherits this lineage. It connects the two, but at the same time keeps them apart. The hardware/software split is one version of the broader machine–user divide. This same manner of separation also shaped the culture of early computing (Levy, 1984; Hirschheim & Klein, 2012).

Stephen Levy's (1984) account of the hacker culture in California describes how source code was often open, free, and without passwords among enthusiastic hobbyists. There were no artificial boundaries, but in one line of thought (Stallman, 2002) everything was about developing systems together, openly, with no passwords or user accounts restricting access. While a distinction existed between developers and the system, they were nonetheless united as part of a shared collective project of development. However, boundaries existed right beside the hackers' ideology. Levy (1984) describes how large companies such as IBM or Xerox hid their code and this caused repulsive reactions among the hackers. Hidden source code created a split between computers and developers, i.e. hackers. For a long time, many wished for no user accounts at all—everything open, nothing hidden—so that users could develop the systems further (Stallman, 2002). This ethos implied no limits and no secrets, while in a related but distinct strand Raymond (1999) emphasized the open, collaborative development model as opposed to proprietary secrecy.

As computing moved into organizations and business systems, the emphasis shifted: proprietary control and secrecy became central. The shift also marked a broader historical transformation in the Information Systems field, as computing practices moved from hobbyist and technical communities into managerial and organizational domains (Hirschheim & Klein, 2012). While developers in the hacker culture were seen, and wanted to be seen, as "wizards," organizational life in business and management drew clear limits between IT systems. Models such as the Strategic Alignment Model (Henderson & Venkatraman, 1993) created separate pockets for different "actors": business goals, IT goals, organizational structures, and so on. In these framings, the user is positioned at a different level altogether, which implies the split.

The same split of social and technical is visible in cybersecurity frameworks and standards. These documents are powerful because they not only guide practice but also define what counts as security. And they repeat the division. Standards rise from best practices, and once formalized they reinforce them, producing a cycle. It is no surprise that the NIST Cybersecurity Framework (CSF) includes this split: some categories are clearly technical (system configuration, encryption, algorithmic safeguards), while others are explicitly human-focused (security awareness, training) (NIST, 2024). ISO/IEC 27001 provides a similar distinction. Access control is described in technical terms—logon procedures, password management—while separate clauses emphasize human responsibility, especially for privileged users (ISO/IEC, 2022). In both, the discourse of a technical–human divide is embedded in the very structure of the standard.

3. The mix – Illustrative Sociomaterial Practices

Intrusion detection systems (IDS) continuously monitor host or network activity to detect deviations from established patterns, using either signature-based methods or anomaly-based approaches (Satılmış, Akleylek & Tok, 2024). Here, the behavioral side becomes particularly compelling. For example, transferring an unusually large amount of data can be flagged as suspicious simply because it breaks from the established rhythm of activity. However, a sophisticated attacker can perform the same act of exfiltration gradually and imperceptibly—changing behavior—thus hiding within the flow of ordinary system operations**.** By mimicking ordinary traffic, the attacker hides exfiltration in slow, low-and-slow flows to avoid detection. As Liao et al. (2013) note, the central challenge of anomaly detection lies in modeling the "normal" so precisely that even subtle infiltration leaves detectable traces. This interplay evokes a form of mimicry, where an intrusion succeeds precisely because it appears ordinary. Today's AI technologies seek to spot anomalies through deep learning algorithms (Naseer et al., 2018).

Yet mimicry is also used in an entirely social manner in technological attacks. A Trojan attack, for instance, depends not only on code but also on trust—it enters the system disguised as something benign, something that a user desires or welcomes. As Bowles and Hernández-Castro (2015) recount in their historical overview of the "Trojan Horse" defense, malware often exploits human expectations of safety and routine to infiltrate systems unnoticed. Trojan attacks exploit trust by disguising malicious elements as benign artefacts. Such a deception took place and had consequences, as illustrated by the 2024 incident in which pagers used by Hezbollah were covertly weaponized through pre-installed explosive components—a Trojan horse, explosives hidden within an everyday communication device (BBC, 2024). The pagers, which were considered to be safer than mobile phones, were compromised. In cybersecurity, the user desires, conventions, and habits that are considered social are mixed deeply with technology (Vuorinen & Tetri, 2016). Thus, as there is desire to use something, there is also a possibility to exploit this desire by offering Trojan horses. However, the same logic extends to defense, where decoys are used to attract and misdirect attackers (e.g., Bowen et al., 2009).

An intrusion detection system defining "normal" behavior or a Trojan attack exploiting trust both show how deeply the social and technical are fused. The same applies to cybersecurity processes such as risk assessment, where protection depends on the perceived value of assets (ISO/IEC, 2022; Baskerville, 1993). Different assets receive different levels of defense, a logic that appears rational but is grounded in collective judgment: what is worth protecting and why. These values do not arise from technology itself but from socio-technical negotiation—how assets acquire meaning within business goals and how technical means enable or constrain that meaning. In this sense, value and security alike emerge from the interplay between social intention and technical possibility.

This means that security decisions are always contextual, situated within what is wanted (the social) and what can be done (the technical). The modalities of these activities—possibility, limitation, and potential—constantly mix the two. Technology is not mute here; it actively participates in shaping security decisions through its capabilities and constraints. A good example is the case of quantum technologies: they are not yet

a direct threat, but they have the potential to become one. Therefore, today's encryption must be designed with future quantum capabilities in mind. This requirement brings together multiple layers: the actual (what the situation currently is), the potential (what it might become), the technical (quantum computation), and the social (the desire to maintain confidentiality now and in the future). Recent research also highlights how advances in quantum and supercomputing are already reshaping cybersecurity strategy and preparedness (Yalcin et al., 2024).

There is no way to extract one aspect without bringing a piece of another with it. In this sense, security cannot be understood purely socially or purely technologically. I argue that this applies to every dimension of cybersecurity. The two are always mixed—each carrying traces of the other. Security itself is a social manifestation that always depends on material and technological dimensions.

4. Overcoming the Split? Implications for Research and Practice

This section outlines two paths—one in research and one in practice—for challenging the split. In research, the key question is how to treat the technical and the social symmetrically. As shown above, the field remains bifurcated: we have theories that focus on the technological side and others that focus on the human side, usually framing people either as problems to be managed or, through awareness and training, as "human sensors" who assist in security tasks (Vielberth et al., 2019). Very few approaches attempt to integrate both dimensions. The sociomaterial turn in Information Systems has been one major attempt to do so, emphasizing the mutual shaping of technology and human activity (Orlikowski, 2007; Orlikowski & Scott, 2008).

In terms of research, when it comes to overcoming the gap, sociomaterialism is certainly an option and, without doubt, provides a wide range of usable concepts that help to view technological and human agency symmetrically. However, the difficulty with this approach is that sociomaterialism does not offer a single, straightforward way to approach a subject, to inquire into knowledge, or to conduct analysis (Cecez-Kecmanovic et al., 2014), even though there are more structured methodological attempts (Gaskin et al., 2014). While it must be admitted that reality is complex and difficult to understand—and it is precisely here that the human–technology split has helped us simplify contextual complexity—it would be useful if there were a simpler approach, one that already has a strong research tradition.

Such an opportunity is provided by ANT, which was developed within science and technology studies in the late 1970s and early 1980s as a reaction to how narrowly agency had been understood—as something that belonged only to humans. Material objects were typically treated as passive tools or background props, while humans were seen as the sole decision-makers and sources of action. This is the same discussion that later emerged in the Information Systems field. Latour and Woolgar (1979) in Laboratory Life showed how scientific instruments are not neutral intermediaries but crucial participants in knowledge production. Michel Callon (1986) likewise demonstrated how technological artifacts can act, translate, and enroll others in networks of relations. This may not sound revolutionary, but it breaks a long Western philosophical lineage that placed the human at the center of agency. ANT, by contrast, treats human and material actors symmetrically. Despite its name, ANT is arguably more of a methodology than a theory, and this is precisely its strength.

The dynamic core of ANT is that actors are defined not by their intrinsic properties but by their effects on others—meaning that anything that has an effect on its surroundings is an actor (Callon & Latour, 1981). For example, a chair, which I am sitting on, enables writing; if removed, its effects become visible. The chair is not defined by its inner properties, but by the effects it produces—it makes sitting possible. Similarly, each new connection in a network changes the actor-network slightly, and thus the network evolves as relations emerge and disappear. When a new actor becomes part of a security assemblage, the configuration changes. If a new user joins a system, it may be a small adjustment overall—a mere scaling up—but still a change in the relations of trust, control, and vulnerability. Each new actor, human or nonhuman, has effects: one more potential attack vector, one more dependency, one more mediator. If the system instead connects to malicious code, the shift can be more dramatic. In either case, ANT invites us to see these changes symmetrically—as transformations in a network of relations rather than as events confined to a "human" or "technical" side. In this way, ANT provides a framework that closes, or at least narrows, the human-technology gap.

In security architectures, standards, and frameworks—as discussed above—the treatment typically follows the split. There are specialists for both sides: technology experts focusing on system configurations, and behavioral specialists focusing on human factors. However, not every development in the field adheres to this division at the level of principle. The Zero Trust security architecture, which has gained significant attention in recent years, offers a notable example. It aligns well with contemporary working models such as bring your own

device (BYOD) and the distributed nature of modern organizations. Unlike perimeter-based models, Zero Trust does not rely on a predefined boundary that separates trusted insiders from untrusted outsiders. Instead, it responds to the realities of networked, hybrid environments where such boundaries are fluid or nonexistent.

Zero Trust is distinctive because it closes the gap through its principle of continuous verification—summarized by NIST as the maxim never trust, always verify (Rose et al., 2020). In the Zero Trust framework, no entity—human or technical—is trusted by default. Every actor, whether a user, device, or service, must continually prove its legitimacy through authentication and authorization processes. If a user begins to perform actions that deviate from established patterns, the system flags this as potentially suspicious. Likewise, if a device suddenly changes its location or access behavior, it is subjected to additional checks. In Zero Trust, trust is not a permanent status but a temporary and revocable condition, granted only for as long as the verification holds.

Through this logic of continuous distrust, Zero Trust effectively treats human and technical entities in a symmetrical way. Both are considered active participants within the security assemblage, capable of legitimate or illegitimate actions. The system's role is not to judge intent but to verify behavior. By subjecting users and devices alike to the same rule of conditional trust, Zero Trust embodies a practical realization of the symmetry that ANT describes at a conceptual level. It demonstrates how, in practice, the boundary between social and technical can be narrowed through concrete design principles rather than rhetorical redefinition.

5. Discussion and Conclusions

The split between the social and the technical is an old one. It extends far beyond the fields of Information Systems and cybersecurity and is, in fact, an overarching feature of Western thought rather than a local peculiarity. Yet, although the split may feel intuitively functional—"humans are different from machines, right?"—it can also create a haze, an illusion of two distinct worlds. This illusion divides specialists into two camps: those who "know the technical side" and those who "understand human behavior." However, if we look more closely at how systems operate and how they are established, these two sides are constantly intertwined. There is no purely technological, nor purely social. What we have instead is a continuous mixture—something we can begin to approach differently through frameworks such as Zero Trust and ANT. Importantly, this argument does not deny that contemporary cybersecurity practice routinely combines technical controls with training, policies, and procedures; rather, it questions the conceptual framing that continues to treat these elements as ontologically distinct domains.

The practical takeaway of this paper could be summarized as follows: whenever a practitioner thinks that something belongs to the "human side" or the "technical side," that moment should trigger a pause. It is an opportunity to ask whether there is something social embedded within the technical—and vice versa. What social and technological ramifications arise across the entire security assemblage when one element is altered? How do people think about the devices and technologies they engage with, and how do these beliefs influence their actual behavior? The social is not confined to people, and the technical is not merely a platform on which we act; it also shapes how we think, perceive, and decide. Could a situation be framed differently? Could technologies be used in another way? What principles or values does the system ultimately serve? These are the questions that help us to move beyond the split. They are not silver bullets—nothing in this field is—but they can help practitioners to see the broader picture and to recognize how both sides continually fold into each other.

On the research side, the field does not necessarily need to be reorganized from scratch, but it should be re-imagined as operating on a single, shared level. ANT may assist in this task. It is not a magical new lens that reveals everything at once, but rather a subtle shift: a way of understanding that, in research as in practice, we might not need to analytically separate the technical and the social in organizational contexts, where they constantly co-produce one another.

While contemporary cybersecurity research and practice increasingly emphasize holistic approaches that combine technical, human, and organizational considerations, these efforts often remain framed through analytically distinct categories. Humans and technologies are coordinated, yet still treated as different kinds of entities with separate explanatory roles. This paper does not deny such integrative efforts, but questions the persistence of this underlying separation and proposes an alternative way of thinking about agency, responsibility, and trust in cybersecurity. By approaching security as a field of heterogeneous but symmetrical actors, it becomes possible to reflect differently on how controls, behaviors, and infrastructures jointly produce security outcomes. From this perspective, Zero Trust can be understood not merely as a technical

architecture, but as a practical instantiation of a broader conceptual shift with implications for both research and practice.

AI Declaration: Portions of this paper were refined with the assistance of OpenAI's ChatGPT. The tool was used to support language refinement, grammatical editing, and reference formatting. All ideas, interpretations, argumentation, and substantive content are the author's own.

Ethics declaration: No specific ethical clearance was needed for this theoretical research.

References

Adams, A. and Sasse, M.A. (1999) "Users Are Not the Enemy", *Communications of the ACM*, Vol. 42, No. 12, pp. 40–46.

Baskerville, R. (1993) "Information Systems Security Design Methods: Implications for Information Systems Development", *ACM Computing Surveys (CSUR)*, Vol. 25, No. 4, pp. 375–414.

BBC News (2024) *Israel Says Hezbollah Pager Blasts Were Pre-Planned Trojan Attack*, 17 September. [online] Available at: https://www.bbc.com/news/articles/cwy3l02wxqdo (Accessed: 2 October 2025).

Bowen, B.M., Hershkop, S., Keromytis, A.D. and Stolfo, S.J. (2009) "Baiting Inside Attackers Using Decoy Documents", in *International Conference on Security and Privacy in Communication Systems*, Berlin and Heidelberg, Springer, pp. 51–70.

Bowles, S. and Hernandez-Castro, J. (2015) "The First 10 Years of the Trojan Horse Defence", *Computer Fraud & Security*, Vol. 2015, No. 1, pp. 5–13.

Bulgurcu, B., Cavusoglu, H. and Benbasat, I. (2010) "Information Security Policy Compliance: An Empirical Study of Rationality-Based Beliefs and Information Security Awareness", *MIS Quarterly*, Vol. 34, No. 3, pp. 523–548.

Callon, M. (1986) "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay", in J. Law (ed.) *Power, Action and Belief: A New Sociology of Knowledge?* London, Routledge, pp. 196–233.

Callon, M. and Latour, B. (1981) "Unscrewing the Big Leviathan: How Actors Macro-Structure Reality and How Sociologists Help Them to Do So", in K. Knorr-Cetina and A.V. Cicourel (eds.) *Advances in Social Theory and Methodology: Toward an Integration of Micro- and Macro-Sociologies*. Boston, MA, Routledge & Kegan Paul, pp. 277–303.

Cecez-Kecmanovic, D., Galliers, R.D., Henfridsson, O., Newell, S. and Vidgen, R. (2014) "The Sociomateriality of Information Systems", *MIS Quarterly*, Vol. 38, No. 3, pp. 809–830.

Cybersecurity and Infrastructure Security Agency (CISA) (2024) *Widespread IT Outage Due to CrowdStrike Update*. [online] Available at: https://www.cisa.gov/news-events/alerts/2024/07/19/widespread-it-outage-due-crowdstrike-update (Accessed: 2 October 2025).

CrowdStrike (2024a) *Falcon Sensor Update Incident — Technical Summary*, CrowdStrike Blog, 19 July. [online] Available at: https://www.crowdstrike.com/ (Accessed: 2 October 2025).

CrowdStrike (2024b) *Guidance for Customers Remediating Channel File 291*, CrowdStrike Support Advisory, 20 July. [online] Available at: https://www.crowdstrike.com/ (Accessed: 2 October 2025).

Foucault, M. (1970) *The Order of Things: An Archaeology of the Human Sciences*. London, Tavistock.

Gaskin, J., Berente, N., Lyytinen, K. and Yoo, Y. (2014) "Toward Generalizable Sociomaterial Inquiry", *MIS Quarterly*, Vol. 38, No. 3, pp. 849–872.

Henderson, J.C. and Venkatraman, N. (1993) "Strategic Alignment: Leveraging Information Technology for Transforming Organizations", *IBM Systems Journal*, Vol. 32, No. 1, pp. 4–16.

Hirschheim, R. and Klein, H.K. (2012) "A Glorious and Not-So-Short History of the Information Systems Field", *Journal of the Association for Information Systems*, Vol. 13, No. 4, pp. 188–235.

International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) (2022) *ISO/IEC 27001:2022 Information Security, Cybersecurity and Privacy Protection — Information Security Management Systems — Requirements*. Geneva, ISO.

Latour, B. (1993) *We Have Never Been Modern*. Cambridge, MA, Harvard University Press.

Latour, B. (2005) *Reassembling the Social: An Introduction to Actor-Network Theory*. Oxford, Oxford University Press.

Latour, B. and Woolgar, S. (1979) *Laboratory Life: The Construction of Scientific Facts*. Beverly Hills, CA, Sage.

Levy, S. (1984) *Hackers: Heroes of the Computer Revolution*. New York, Anchor.

Liao, H.-J., Lin, C.-H.R., Lin, Y.-C. and Tung, K.-Y. (2013) "Intrusion Detection System: A Comprehensive Review", *Journal of Network and Computer Applications*, Vol. 36, No. 1, pp. 16–24.

Loch, K.D., Carr, H.H. and Warkentin, M.E. (1992) "Threats to Information Systems: Today's Reality, Yesterday's Understanding", *MIS Quarterly*, Vol. 16, No. 2, pp. 173–186.

Naseer, S., Saleem, Y., Khalid, S., Bashir, M.K., Han, J., Iqbal, M.M. and Han, K. (2018) "Enhanced Network Anomaly Detection Based on Deep Neural Networks", *IEEE Access*, Vol. 6, pp. 48231–48246.

Nohlberg, M. (2009) "Why Humans Are the Weakest Link", in M. Gupta and R. Sharman (eds.) *Social and Human Elements of Information Security: Emerging Trends and Countermeasures*. Hershey, PA, Information Science Reference, pp. 15–26.

Orlikowski, W.J. (2007) "Sociomaterial Practices: Exploring Technology at Work", *Organization Studies*, Vol. 28, No. 9, pp. 1435–1448.

Orlikowski, W.J. and Scott, S.V. (2008) “Sociomateriality: Challenging the Separation of Technology, Work and Organization”, *Academy of Management Annals*, Vol. 2, No. 1, pp. 433–474.

Raymond, E.S. (1999) *The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary*. Sebastopol, CA, O’Reilly.

Rose, S., Borchert, O., Mitchell, S. and Connelly, S. (2020) *Zero Trust Architecture (NIST Special Publication 800-207)*. Gaithersburg, MD, National Institute of Standards and Technology. [online] Available at: https://doi.org/10.6028/NIST.SP.800-207(Accessed: 2 October 2025).

Satılmış, H., Akleylek, S. and Tok, Z.Y. (2024) “A Systematic Literature Review on Host-Based Intrusion Detection Systems”, *IEEE Access*, Vol. 12, pp. 27237–27266.

Siponen, M., Baskerville, R. and Heikka, J. (2006) “A Design Theory for Secure Information Systems Design Methods”, *Journal of the Association for Information Systems*, Vol. 7, No. 11, pp. 725–770.

Stallman, R.M. (2002) “On Hacking”, in P. Himanen (ed.) *The Hacker Ethic and the Spirit of the Information Age*. New York, Random House, pp. 3–9.

Vielberth, M., Böhm, F., Fritsch, L. and Pernul, G. (2019) “Security Operations Center: A Systematic Study and Open Challenges”, *Computers & Security*, Vol. 87, p. 101579. https://doi.org/10.1016/j.cose.2019.101579

Vuorinen, J., & Tetri, P. (2016). “Paradoxes in information security”, *IEEE Potentials,* 35(5), 36-39.

Legal Protection against Digital Personal Identity Fraud in South Africa

Murdoch Watney
University of Johannesburg, South Africa

mwatney@uj.ac.za

Abstract: As digital transformation accelerates worldwide, personal identity authentication has become a global issue. In South Africa digital personal identity has become central to how individuals interact with government services, financial institutions, and online platforms. Digital identity refers to the authentication of users through personal attributes such as biometric data, voice, image, and behavioural patterns. While these systems offer efficiency and accessibility, they also introduce significant threats. Increasingly, digital personal identity is being cloned, manipulated, or misused through techniques like deepfakes, voice cloning, and biometric spoofing. The discussion examines whether South African law adequately protects individuals against the misuse of digital personal identity, particularly in the context of fraud. While financial gain is a common motive, misuse may also result in reputational harm, privacy violations, and unauthorised commercial exploitation. The analysis considers the legal framework governing digital personality fraud under the Protection of Personal Information Act (POPIA), the Cybercrimes Act, the Consumer Protection Act (CPA), and the common law remedy of *actio iniuriarum*, within the framework of constitutional rights to dignity and privacy. Comparative insights from Denmark, Tennessee (United States), the European Union, and the United Kingdom help contextualise South Africa's approach. The discussion argues that while South African law provides substantive protection, enforcement remains challenging due to technological complexity, evidentiary burdens, and limited institutional capacity. Strengthening technical expertise, public awareness, and regulatory enforcement is essential to ensure meaningful protection for digital citizens. Drawing on global approaches, the discussion proposes targeted legal reforms to enhance accountability and safeguard digital personal identity against fraud.

Keywords: Digital personal identity, Personality rights, Digital identity fraud, Digital personal identity authentication, South African legal framework governing digital identity fraud

1. Introduction

South Africa is rapidly digitising its public and private sectors. Citizens increasingly access government services, financial platforms, and personal technologies through biometric and digital authentication systems. This transformation has elevated digital personal identity which is the electronic representation of unique traits such as fingerprints, voice, and facial features as a cornerstone of modern life. Yet, as authentication becomes more seamless, the risks of personal identity misuse grow. Technologies like deepfakes, voice cloning, and biometric spoofing have enabled sophisticated forms of impersonation, often resulting in fraud, reputational harm, or privacy violations.

These abuses raise urgent questions about the adequacy of South African law in protecting digital citizens. While cybersecurity frameworks address technical, legal, and human behavioural components, they may not be sufficient to protect digital personal identity. This is because digital identity misuse, especially through AI-driven impersonation, raises deeper legal questions about autonomy, dignity, and expressive control.

Despite its growing role in authentication and fraud prevention, digital personal identity has not received the focused legal attention it demands. The legal system must respond not only to cybersecurity threats, but also to infringements of personality rights, which protect the core attributes of personhood. Watney's (2024) analysis of cyber fraud within the South African cybersecurity legal framework, focused primarily on the criminalisation of online deception and data breaches. However, the specific vulnerability of digital personal identity as a legally protected facet of personhood remains underexplored. The discussion builds on that foundation by examining fraud through the lens of personal identity misuse and evaluating the adequacy of legal protection for digital citizens.

The discussion examines whether South African law provides effective protection against fraud arising from the misuse of digital identity, and considers how global models, such as Denmark's proposed likeness-as-property regime and the Tennessee ELVIS Act, may inform potential reform.

2. Background to the Discussion: Defining key Concepts

Before evaluating the legal framework, it is essential to define the key concepts that underpin digital identity and its misuse.

2.1 Personality Rights vs Personal Identity

Personality rights in South African law protect the legal rights that arise from being a person such as dignity, privacy, identity, and reputation. These rights are afforded to individuals over aspects of their personal identity that express their personhood such as their name, image, voice, likeness, and reputation (Neethling, 2005). They protect:

- Image from being used without consent;
- Voice from being cloned;
- Reputation from being harmed; and
- Privacy from being invaded.

In this context, personal identity is not a separate legal category, but rather the content protected by personality rights. These rights are rooted in Roman-Dutch common law, where they were traditionally applied to protect individuals in the physical world from unauthorised use or exploitation of their identity (Neethling, 2005). The court in *Grütter v Lombard* confirmed the existence of personality rights in common law and found that the unauthorised use of a person's image in advertising infringed his right to dignity and identity. This case illustrates how the misuse of personal identity such as one's image or likeness constitutes a violation of personality rights.

2.2 Digital Personal Identity

With the rise of digital technologies, these rights have migrated online.

There is not a single or general definition of digital identity (Robles-Carrillo, 2024). Digital identity refers to the data and attributes used to authenticate an individual online. The authentication data consists of biometric data such as fingerprints and facial recognition, voice and image, as well as personal identifiers such as ID numbers, and behavioural patterns (Robles-Carrillo, 2024). These elements are increasingly used to verify identity in banking, healthcare, education, and government services. However, when cloned or manipulated, they can be weaponised to commit fraud.

A person's identity is not only expressed through physical presence but also through digital representations including biometric data, voice recordings, photographs, and even AI-generated likenesses. Just as a person's image could be misused in a newspaper advertisement decades ago, it can now be cloned, manipulated, and distributed across digital platforms without consent. However, drawing an analogue between the physical and electronic mediums is complex, as the electronic medium presents challenges that do not exist in the physical realm. Some of these challenges are:

- Security and privacy: Unlike physical documents, electronic data can be intercepted, duplicated, or altered remotely. Ensuring confidentiality and integrity requires for example sophisticated encryption and access controls.
- Authenticity and trust: In the physical world, signatures, seals, and paper trails help verify authenticity. Authenticity online is established through digital signatures, certificates, and blockchain-like technologies which are powerful but not always easily understood.
- Permanence vs. ephemerality: Physical media often have a tangible permanence. Electronic data, while easily duplicated, can be deleted, corrupted, or lost due to system failures or cyberattacks.

2.3 Misuse of Digital Identity

Misuse of digital identity refers to the unauthorised or deceptive use of a person's digital attributes such as biometric data, voice, image, or personal identifiers to impersonate, manipulate, or defraud others, often resulting in financial, reputational, or privacy-related harm. This misuse may occur through technological manipulation, such as deepfakes and biometric spoofing, or through human manipulation, such as social engineering.

The online environment amplifies the risk of misuse**.** Digital content is easily copied, altered, and disseminated at scale, often without the subject's knowledge. Technologies such as deepfakes and voice cloning make it possible to impersonate someone with alarming realism, leading to reputational harm, privacy violations, and even financial fraud.

Personality rights must therefore be understood as continuing protections that apply both offline and online. Although the legal principles remain the same, the medium and magnitude of harm differ. In the digital age, the

misuse of personality rights is not only easier but also more difficult to detect and remedy, raising urgent questions about enforcement, consent, and accountability.

2.4 Cybersecurity Threats and Risks

There is not a universal definition for cybersecurity. Cybersecurity is defined as "the practice of protecting systems, networks and data from digital threats. It involves strategies, tools and frameworks designed to safeguard sensitive information and ensure the integrity of digital operations" (Yassar et al, 2025). The cybersecurity legal landscape is shaped by the imperative to identify digital threats and assess the likelihood of those threats materialising into harm.

Effective cybersecurity requires a holistic framework that integrates three interdependent components: technical**,** legal, and human. The technical component encompasses encryption, firewalls, biometric authentication, and intrusion detection systems designed to prevent unauthorised access and data breaches. The legal component involves statutory and regulatory instruments that establish rights, duties, and enforcement mechanisms for data protection and cyber resilience. The human component addresses behavioural vulnerabilities, including social engineering, insider threats, and the need for digital literacy and ethical awareness. A robust cybersecurity framework must account for all three dimensions, recognising that technological safeguards alone are insufficient without legal accountability and informed human conduct. In South Africa, this tripartite approach is essential to mitigate the misuse of digital identities and ensure institutional and individual resilience against evolving cyber threats.

In the context of digital identity, key threats include biometric spoofing, deepfake impersonation, voice cloning used in scams, and data breaches exposing identity profiles. These threats are exacerbated by the permanence of biometric data as unlike passwords, fingerprints cannot be changed once compromised.

2.5 Cybercrime and Cyber Fraud

The South African National Cybersecurity Policy Framework (NCPF) defines cybercrime as: "Criminal and other offences that are committed through the use of cyberspace or information and communication technologies (ICTs), including the internet, mobile devices, and computer systems". Under the NCPF cybercrime explicitly includes the misuse of digital identities.

Section 8 of the Cybercrimes Act defines cyber fraud "when a person unlawfully and with the intent to defraud makes a misrepresentation by means of data or a computer program, or through any interference with data or a computer program, which causes actual or potential prejudice to another".

Digital identity misuse enables various forms of fraud, such as:

- Impersonation

This involves using deepfakes, voice clones, or stolen credentials to pose as someone else. For example, a cloned voice may be used to authorise a bank transfer, or a deepfake video may impersonate a CEO to manipulate stock prices. Deepfake technology uses artificial intelligence (AI) to manipulate audio and visual data. It can mimic voices, replicate faces, and fabricate entire personas. This enables the creation of synthetic versions of individuals without their knowledge or consent, deceiving the public and undermining trust (Mashinini, 2025). AI-generated deepfake challenges core legal principles such as consent, privacy, intellectual property, and criminal liability

- Biometric spoofing

Biometric spoofing refers to faking fingerprints, facial scans, or iris patterns to gain access to devices or systems. This form of fraud is particularly serious because biometric data is immutable.

- Social engineering

Social engineering involves manipulating individuals into revealing sensitive information or performing actions. One example is Business Email Compromise (BEC), where the threat actor impersonates executives or vendors to trick employees into transferring funds. Deepfakes and voice cloning are increasingly used to enhance credibility in these scams.

- Data manipulation

This includes altering or fabricating identity data to deceive systems, such as creating fake ID profiles or modifying biometric records.

3. South African Legal Framework Governing Digital Identity Fraud

Having defined key concepts**,** the South African legal framework governing digital personal identity fraud is explored.

It is important to distinguish between general legislation such as the Protection of Personal Information Act 4 of 2013 (POPIA), the Consumer Protection Act 68 of 1998 (CPA), and the Cybercrimes Act and industry-specific regulation**,** namely the Joint Standard 2 of 2024 (Joint Standard 2)**.**

POPIA, CPA, and the Cybercrimes Act apply broadly across sectors and are designed to protect all individuals in South Africa, regardless of the context in which their personal information is processed or their consumer rights are engaged. These statutes establish foundational rights and obligations relating to privacy, data protection, fraud prevention, and fair commercial conduct. By contrast, the Joint Standard 2 applies specifically to financial institutions regulated under the Financial Sector Regulation Act. It imposes cybersecurity and cyber resilience obligations tailored to the operational risks of the financial sector. While it complements POPIA by reinforcing the duty to protect biometric and identity-related data, its scope is limited to regulated entities such as banks, insurers, and pension funds. It does not create general rights for individuals, but rather sets technical and governance standards for institutional compliance.

3.1 Constitutional Rights

The Constitution guarantees the right to dignity in section 10, privacy in section 14, and bodily integrity in section 12. These rights underpin the legal protection of digital identity. Unauthorised use of a person's image or voice may violate these rights, especially when used to deceive or defraud.

South Africa's commitment to the United Nations Sustainable Development Goals (UN SDG) with specific reference to UN SDG 16 further supports the need for strong legal frameworks to protect individuals from exploitation and promote justice and accountability.

3.2 Protection of Personal Information (POPIA)

POPIA applies to any person or entity that processes personal information in South Africa, across all sectors. It gives effect to the constitutional right to privacy by regulating how personal information which includes identity-related data is collected, stored, and used. POPIA is therefore about data governance and privacy. Under POPIA and the EU GDPR, biometric data is classified as special personal information because of its sensitivity and potential for misuse

POPIA defines personal information broadly to include:

- Biometric data;
- Photographs;
- Voice recordings;
- Identity numbers; and
- Any data that can identify a person directly or indirectly.

Under POPIA:

- Consent is required before processing personal information (section 11);
- Purpose limitation ensures data is used only for its intended function (section 13); and
- Security safeguards must be in place to prevent unauthorised access or misuse (section 19).

In digital identity misuse, POPIA is triggered when:

- A person's biometric or identity data is cloned, shared, or manipulated without consent; and
- A company fails to secure personal data.

When someone's image, voice, or biometric data is cloned or manipulated without consent, it may constitute both a violation of personality rights and unlawful processing under POPIA.

3.3 Consumer Protection Act (CPA)

The CPA applies to commercial transactions where a person buys, uses, or is affected by a product or service. It protects consumers from unfair, misleading, or deceptive conduct in the supply of goods and services. CPA is therefore about market fairness and consumer rights.

In digital identity misuse, CPA is triggered when:

- A cloned voice is used to authorise a fraudulent purchase.
- A consumer is misled by a fake endorsement or impersonation in advertising.
- A service provider misuses identity data in a way that affects the consumer's transaction.

3.4 Cybercrimes Act

The Cybercrimes Act criminalises unlawful access to data, cyber fraud, identity theft, and cyber extortion. It provides a framework for prosecuting digital identity misuse, especially when biometric data or cloned media is used to deceive systems or individuals (Watney, 2024).

3.5 Joint Standard on Cybersecurity and Cyber Resilience(Joint Standard 2)

The Joint Standard 2 was issued by the Financial Sector Conduct Authority (FSCA) and the Prudential Authority (PA)**,** this standard requires financial institutions to implement robust cybersecurity measures. While not directly focused on digital identity, it reinforces the duty to protect biometric and personal data used in authentication.

In conclusion: While both POPIA and the CPA may be relevant in cases of digital identity misuse, they operate in distinct legal domains. POPIA governs the lawful processing and protection of personal information, focusing on consent, privacy, and data security. The CPA, by contrast, applies to transactions between suppliers and consumers in a commercial context. In fraud scenarios involving digital identity, POPIA may be invoked where personal data is misused without consent, while the CPA applies where such misuse results in consumer harm within a business context.

The Joint Standard 2 plays a critical role in regulating the security of digital identity systems, particularly where biometric authentication is used. It requires financial institutions to implement governance frameworks, risk assessments, and incident response plans to prevent unauthorised access, data breaches, and identity-based fraud. In cases where a financial institution fails to secure biometric login systems or allows impersonation through cloned voice or image data, it may be in breach of both POPIA and the Joint Standard 2. The Standard 2 complements POPIA by reinforcing the technical and organisational safeguards required to protect personal information, and it intersects with the CPA when identity misuse results in consumer harm during financial transactions. The Joint Standard 2 therefore forms part of a layered regulatory approach to digital identity protection, especially in high-risk sectors where fraud and impersonation are prevalent.

4. Legal Liability and Enforcement Challenges Under the South African law

Having outlined the legal framework governing digital identity protection in paragraph 4, the following section examines its practical application regarding liability and the challenges associated with enforcing liability.

4.1 Criminal Liability

Digital identity misuse may constitute a cybercrime under the Cybercrimes Act which criminalises offences such as identity theft, cyber fraud, and unauthorised access to data. These offences require proof of intent to deceive and actual or potential prejudice. However, prosecution is often hindered by practical challenges. Perpetrators frequently operate anonymously or across borders, using encrypted platforms, spoofed IP addresses, and false digital credentials, making them difficult to trace. Even when a suspect is identified, gathering admissible digital evidence demands specialised forensic expertise and strict chain-of-custody protocols.

Law enforcement agencies may lack the technical capacity or resources to investigate complex cyber offences, particularly those involving biometric spoofing or deepfake impersonation. In addition, prosecutorial discretion and systemic backlog in the criminal justice system may result in delays or de-prioritisation of cybercrime cases especially where the financial loss is modest or the victim is an individual. While the legal framework exists, the practical enforcement of criminal liability remains uneven. These challenges underscore the need for enhanced digital forensic capacity, cross-border cooperation, and public-private partnerships to strengthen cybercrime prosecution in South Africa.

4.2 Civil Liability

Where biometric data is processed without consent, victims may seek redress under the POPIA. Available remedies include:

- Lodging a complaint with the Information Regulator.
- Seeking civil damages under section 99.

- Requesting correction or deletion of unlawfully processed data.

Victims may also pursue common law claims for infringement of personality rights, including unauthorised use of image or voice, defamation, invasion of privacy, or misrepresentation. These claims are typically brought through the *actio iniuriarum,* a delictual action that protects non-patrimonial interests such as dignity, identity, and autonomy. Damages may encompass financial loss, emotional distress, and reputational harm.

Despite these remedies, enforcement remains fraught with difficulty. Identifying perpetrators in cyber-related cases is notoriously complex, particularly when anonymity, spoofing, or foreign servers are involved. In some instances, victims may consider holding third parties such as banks or service providers liable for failing to implement adequate safeguards, especially where biometric authentication systems are compromised. Yet this route is equally challenging as proving negligence or breach of statutory duty requires expert evidence, and institutions may rely on contractual disclaimers or demonstrate compliance with regulatory standards.

Moreover, civil litigation in South Africa is time-consuming and costly. Thus, although the legal framework offers substantive remedies, the real-world enforcement of civil liability remains uncertain particularly for ordinary individuals without access to legal resources or technical expertise.

Against the backdrop of substantive legal protection and practical enforcement challenges, the following section examines how selected foreign jurisdictions have approached the regulation of digital identity misuse, offering comparative insights that may inform South Africa's evolving legal response to digital personal identity fraud.

5. Comparative Analysis

While this discussion focuses on South Africa's legal framework, the challenges of digital personal identity misuse are global in scope. Jurisdictions across the globe are grappling with the legal implications of digital identity misuse, particularly in the context of artificial intelligence and biometric technologies. In light of the enforcement challenges facing South African law, it is instructive to examine how other jurisdictions have responded to the misuse of digital personal identity.

5.1 Global Approaches to Digital Identity Protection

Global approaches to digital identity protection reflect distinct legal philosophies. The UK's *Economic Crime and Corporate Transparency Act* offers a valuable institutional liability approach that merits consideration in South Africa's context. It's*"* failure to prevent fraud" (FTPFP) offence shifts the emphasis from the method of identity misuse to the responsibility of organisations to implement reasonable fraud prevention measures. The FTPF offence is not about digital identity authentication *per se* but about corporate accountability for fraud. This approach is particularly relevant in South Africa's context, where digital identity fraud often occurs within institutional systems that lack robust safeguards.

While the approaches vary, a comparative analysis reveals the following models:

- Statutory personality rights model

The Ensuring Likeness Voice and Image Security (ELVIS) Act, passed in 2024 by Tennessee, US is an example of a statutory personality rights model. It is the first-of-its-kind state law offering protections for an individual's voice and likeness against unauthorized clones and fakes generated using artificial intelligence (AI) (Auerbach, 2025).

The ELVIS Act represents a targeted statutory intervention aimed primarily at performers and public figures against commercial exploitation (Auerbach, 2025). It criminalises and civilly penalises the unauthorised use of a person's voice and likeness, including AI-generated impersonations. Although limited in geographic scope and focused on commercial exploitation, the Act affirms the principle that digital identity misuse constitutes a legal harm. It offers a statutory remedy rooted in personality rights, reinforcing the idea that individuals should control how their identity is used in digital spaces.

- Intellectual property model

Denmark proposes a bill that will grant its citizens copyright to their face, body, and voice. The proposed bill is the first in Europe to give citizens the legal right to ask platforms like Facebook, Instagram, and TikTok to take down fake digital versions of themselves online (Podadera, 2025; Woods, 2025).

The aim of the proposed bill is to:

- Give individuals ownership over their face, voice, and expressions.

- Allow them to license or restrict use like a copyrighted work.
- Provide legal remedies even when no financial loss is proven.

Denmark's proposed legislation represents a proactive shift in how digital identity is protected. Rather than relying solely on privacy-based frameworks like the EU GDPR which regulate how personal data is processed, Denmark reframes likeness (including voice, face, and expressions) as a proprietary right. This gives individuals affirmative control over their digital personal identity, allowing them to demand removal of deepfake content and sue for damages without needing to prove data harm or breach (Podadera, 2025**)**. It moves beyond enforcement of privacy violations and toward ownership-based protection, where the mere unauthorised use of one's likeness is actionable. This model empowers individuals to act swiftly and decisively against misuse, especially in commercial or reputational contexts.

- Hybrid constitutional–statutory model

South Africa's framework is more layered. The Constitution guarantees the rights to privacy and dignity, which underpins the protection of digital identity. POPIA operationalises these rights by regulating the processing of personal and biometric data, while the Cybercrimes Act criminalises identity theft, fraud, and unauthorised access. However, unlike Denmark's proposed IP model or the ELVIS Act's statutory clarity, South Africa has yet to codify digital personality rights explicitly.

5.2 Concluding Observations

While many jurisdictions possess foundational legal frameworks for data protection and identity rights, the rapid evolution of AI-driven fraud has exposed substantive gaps in how digital identity misuse is regulated. Denmark and Tennessee have responded not merely by strengthening enforcement, but by reimagining the legal basis for personal identity protection.

Denmark's proposal to treat a person's likeness including voice, face, and expressions as a form of intellectual property reflects a shift from privacy-based regulation to proprietary control, offering individuals direct remedies against deepfake misuse (Woods, 2025). Although this approach borrows from copyright logic, it does not equate a voice or likeness with a creative work; rather, it creates a *sui generis* right tailored to digital identity. The Denmark approach is novel but it does not automatically mean that other EU member countries will follow suit. Some EU member states may resist this model, arguing that the GDPR already protects biometric data and that extending copyright-like control to likeness may undermine freedom of expression and complicate enforcement. Denmark's proposal to treat likeness as intellectual property may also raise concerns about legal coherence and overreach.

The tension between technological innovation and legal doctrine reflects a broader challenge, namely safeguarding digital identity without overhauling foundational legal principles. Jurisdictions have responded in varied ways. Denmark proposes a likeness-as-property regime applicable to all individuals, while Tennessee's ELVIS Act focuses on protecting performers from AI-driven impersonation in commercial settings.

These developments suggest that legal innovation, and not just enforcement, may be necessary to meet the challenges posed by emerging technologies. South Africa, while equipped with constitutional and statutory protections, may need to consider whether its current framework sufficiently anticipates the risks of digital identity manipulation and whether clearer codification of personality rights in the digital realm is warranted. The goal is not to follow global trends blindly, but to build a coherent, resilient framework that protects digital citizens in practice.

6. Recommendations

Drawing on these insights, the following recommendations aim to strengthen South Africa's legal and institutional response to digital identity misuse, with a focus on fraud prevention, victim redress, and regulatory reform (Mashinini, 2025):

6.1 Enhance Institutional Capacity

- Invest in specialised digital forensic units within law enforcement to improve investigation and evidence-gathering in cybercrime cases.
- Expand the technical expertise and resources of the Information Regulator to ensure effective enforcement of POPIA, especially in biometric data cases.

Improve Legal Clarity and Accessibility

- Develop statutory guidance or judicial practice notes on the application of personality rights to digital identity misuse, including deepfakes and voice cloning.
- Promote public legal education on digital identity rights and available remedies, especially for vulnerable and digitally active users.

6.2 Strengthen Cross-Border and Public-Private Cooperation

- Establish formal cooperation mechanisms with international cybercrime units and digital platforms to trace perpetrators and preserve evidence across jurisdictions.
- Encourage partnerships between government, financial institutions, and tech companies to develop shared standards for biometric authentication and fraud prevention.

6.3 Consider Legislative Reform

- Explore the feasibility of introducing statutory personality rights for digital identity, drawing on models like the Tennessee ELVIS Act and Denmark's likeness-as-property proposal.
- Review the Cybercrimes Act and POPIA to ensure they adequately address emerging threats such as AI-generated impersonation and biometric spoofing.

6.4 Promote Victim-Centred Remedies

- Simplify access to civil remedies for victims of digital identity misuse, including streamlined procedures for POPIA complaints and small claims for non-patrimonial harm.
- Support alternative dispute resolution mechanisms that are affordable, accessible, and tailored to digital harms.

6.5 Introduce Institutional Liability for Failure to Prevent Fraud

- Drawing on the UK's Economic Crime and Corporate Transparency Act, South Africa should consider introducing a statutory offence for failure to prevent fraud**.** This approach shifts liability to organisations that benefit from fraudulent conduct committed by employees or agents, unless they can demonstrate that reasonable fraud prevention procedures were in place. Such a provision would incentivise banks, service providers, and digital platforms to adopt robust identity verification systems and internal controls, thereby reducing opportunities for digital identity misuse. It would also align South Africa's enforcement strategy with global trends in corporate accountability.

7. Conclusion

Digital personal identity has become a cornerstone of modern life, yet its misuse presents complex legal, ethical, and practical challenges. The focus of the research is not solely on the consequence of digital identity misuse such as fraud but on the protection of digital personal identity as an extension of personality rights.

South Africa's legal framework anchored in constitutional rights and comprising POPIA, the Cybercrimes Act, and common law personality rights offers substantive avenues for protection. The court's decision in *Grütter v Lombard* affirms the right to control one's identity, but adapting these principles to digital threats remains a pressing challenge. Victims of digital personal identity misuse, particularly in cases involving cyberfraud, continue to face significant barriers to redress due to technological complexity, evidentiary burdens, and institutional constraints.

Comparative approaches from Denmark, Tennessee, the UK, and the EU illustrate that legal systems can respond to digital identity threats through varied mechanisms, including proprietary control, statutory personality rights, and institutional accountability. These approaches offer valuable insights for South Africa as it seeks to strengthen its liability frameworks and respond to emerging risks such as AI-driven impersonation.

To ensure meaningful protection for digital citizens against digital personal identity fraud, South Africa should enhance its technical capacity, clarify the legal status of digital personal identity, and consider targeted legislative reforms. Measures such as introducing corporate liability for failure to prevent fraud, and exploring statutory personality rights, could help close enforcement gaps and align South Africa's legal response with global best practices.

Ultimately, protecting digital personal identity is not only a cybersecurity imperative, but it is also part of a global effort to safeguard dignity, identity, and trust in the digital age.

AI Declaration: I acknowledge the use of Microsoft Copilot as an assistive tool during my research. The final work reflects my own analysis, argumentation, and interpretation of the subject matter.

Ethics Statement: No ethical clearance was required.

References

Auerbach, B.C. (2024) "Trailblazing Tennessee Legislation – the ELVIS Act", [online], https://www.outsidegc.com/blog/trailblazing-tennessee-legislation-the-elvis-act.

Consumer Protection Act 68 of 1998.

Constitution of South Africa of 1996.

Cybercrimes Act 19 of 2020.

Economic Crime and Corporate Transparency Act (ECCTA). (2024) https://assets.publishing.service.gov.uk/media/67f8ef1845705eb1a1513f35/Failure+to+Prevent+Fraud+Guidance+-+English+Language+v1.6.pdf.

Ensuring Likeness Voice and Image Security (ELVIS) Act of 2024.

Grütter v Lombard 2007 (4) SA 89 (SCA), [online], https://www.saflii.org/za/cases/ZASCA/2007/2.html.

Joint Standard on Cybersecurity and Cyber Resilience Requirements 2 of 2024

Mashinini, N. (2025) "Deepfakes and South African law: remedies on paper, gaps in practice", [online], https://theconversation.com/deepfakes-and-south-african-law-remedies-on-paper-gaps-in-practice-263850.

National Cybersecurity Policy Framework (NCPF) of 2015.

Neetling, J. (2005) "Personality rights: a comparative overview", *The Comparative and International Law Journal of Southern Africa*, Vol. 38, No. 2 2005**.**

Podadera, S.D. (2025), "The first step towards fighting AI Abuse? Denmark Grants Citizens The Copyright to Their Face, Body and Voice", [online], https://www.remotestaff.com.au/blog/denmark-ai-face-copyright-law/.

Protection of Personal Information Act 4 of 2013.

United Nations Sustainable Development Goals, [online], https://sdgs.un.org/goals.

Robles-Carrillo, M. (2024) "Digital identity: an approach to its nature, concept, and functionalities", [online], https://digibug.ugr.es/bitstream/handle/10481/94868/Robles_Carrillo_Digital%20Identity.pdf?sequence=1&isAllowed=y.

Watney, M.M. (2024**)** "Exploring cyber fraud within the South African cybersecurity legal framework" [online]; https://www.researchgate.net/publication/381651578.

Woods, C. Denmark proposes copyright laws to protect against deepfakes", [online] https://lsj.com.au/articles/denmark-proposes-copyright-laws-to-protect-against-deepfakes/.

Yassar et al. (2025) "What is cybersecurity?", (2025) [online], https://www.techtarget.com/searchsecurity/definition/cybersecurity.

Critical Infrastructure, Industrial Control Systems and Cyber Warfare: Ethical and Anticipated Ethical Issues

Richard Wilson[1] and Noah Donnelly[2]
[1]Department of Philosophy/Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA
[2]Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA

wilson@towson.edu
ndonnel1@students.towson.edu

Abstract: Cyber-attacks on Critical Infrastructure and Industrial Control Systems within nation states are very dangerous. They can damage the services and systems that a country needs to maintain safety and stability. There are reasons why these attacks are so problematic. First, Critical Infrastructure and Industrial Control Systems are very important. They regulate the operation of services that are vital to the public such as power grids, water services, roads, hospitals, and banks. Industrial Control Systems also control how things are made in industries like energy and manufacturing. Critical Infrastructure and Industrial Control Systems play a role in keeping a country running. When these systems are disrupted, the economic paralysis, civilian risks, and military vulnerabilities are very similar to what happens in a war. The effects of the disruption of Critical Infrastructure and Industrial Control Systems have large consequences. The way these attacks occur shows that they can cause a lot of problems and help aggressors achieve their goals. These attacks are analogous to bombings. Instead of using physical weapons, belligerents use computers. Because these attacks on infrastructure systems like power and water can cause real physical damage, stop production, or make water undrinkable, they fail to differentiate between military targets and civilians. The economic paralysis, civilian risks, and military vulnerabilities caused by these attacks on Critical Infrastructure and Industrial Control Systems are a problem. These attacks are like acts of war. We are going to examine IC and ICS attacks from two sides, the technical side and the ethical side. We will use examples like Stuxnet, the Ukraine Power Grid compromises the Colonial Pipeline ransomware event, Triton/Trisis safety breaches the Oldsmar water facility incursion, and the JBS meat processing interference. These examples will show what aggressors are trying to accomplish by attacking industrial control systems, how the attacks affect ICS, and how large these operations are, that that these cyber-attacks are really like warfare against the nation states where industrial control systems are located. The attacks upon Industrial control systems are like acts of war. Moreover, this discussion will explore both the ethical dilemmas and anticipated ethical issues with cyber-attacks on CI and ICS. Furthermore, it addresses the morality of categorizing these cyber-attacks as cyber warfare. As a result, examining these cases will help clarify the greater implications of cyber-attacks on nation state CI and ICS systems for national and global security as well as the ethical frameworks governing cyber engagements.

Keywords: Industrial control systems (ICS), SCADA, Programmable logic controllers (PLC), Safety instrumented systems (SIS), Critical infrastructure, Cyber warfare, Modbus, Firmware analysis

1. Introduction

Modern nations rely on a complicated and invisible network of automated systems called Critical Infrastructure (CI). Laying below the surface of everyday life, behind the light switches, tap water, and gas pumps is a large network of Industrial Control Systems (ICS) and Operational Technology (OT). OT is different than Information Technology (IT) that manages data, OT manages physical systems based on laws of physics. OT spins turbines, mixes chemicals, regulates pressure, and opens valves. These systems operated for decades in isolation with proprietary controls, and they were protected due to their running on a closed system (Stouffer et al., 2015). OT was physically separated from the outside world. The drive for efficiency and the move towards remote monitoring has caused OT and IT to merge, which in effect connects legacy industrial systems to the public internet system. This merger has created a new, paralyzing mode of conflict where lines of code can cause physical destruction.

The inherent vulnerability of these systems' inherent vulnerability now represents a fundamental change in how war can be conducted. Previously to destroy infrastructure, the barrier to entry was higher, requiring explosives and an effective way to deliver and detonate them. Moving to today, and state-sponsored hackers can now achieve the same results by manipulating the logic of a Programmable Logic Controller (PLC) from the comfort of their own home (Perlroth, 2021). This capability creates serious ethical questions with regards to the Principles of Distinction and Proportionality. If a cyber-attack were to target a power grid that is used for both civilian and military purposes, not only would it shut off a radar, but it could also eliminate the power to incubators in hospitals, traffic lights in cities, and produce in stores. This now creates a potential for "Cascading Failure", where a localized attack can spread to sectors connected to a power grid. This means that what may have been intended as a limited strike is now effectively a cyber-attack that affects an entire population.

This paper argues that cyber-attacks on ICS are not just criminal acts or espionage, they are acts of war via digital means that violate International Humanitarian Law (Rid, 2012). Through an analysis of the low-level mechanics, we see that these attacks compromise the established "Purdue Model" of industrial control (Williams, 1992). By examining the range of cyber-attacks including the Enterprise Zone down to the Process zone, we identify the fragility of modern infrastructure. Through forensic analysis of case studies including the molecular level sabotage of Stuxnet, the brute-force method used on the Ukraine power grid, and the supply chain stoppage of Colonial Pipeline and JBS. Through those case studies we can demonstrate that the weaponization of critical infrastructure through digital means and cyber-attacks is the greatest ethical challenge of the 21st century.

2. Technical Issues

Understanding the ethical issues related to the severity of ICS warfare, one needs to understand the low-level systems that make these systems so vulnerable. In ICS the priorities are Safety (physically of employees), Reliability and Availability. The priorities of ICS come at the cost of systems security (Stouffer et al., 2015).

2.1 The Purdue Model and Protocol Vulnerabilities

The architecture of ICS can typically be described by the Purdue Enterprise Reference Architecture (PERA) in which the network is segmented into different levels (Williams, 1992). Level 4 is the enterprise business network, that being IT, levels 0-3 are the OT environment. The OT environment ranges Supervisory Control and Data Acquisition systems at Level 2, all the way to the sensors and actuators at level 0. The vulnerabilities of the system lay in the communication protocols between levels. Since OT includes a variety of legacy systems, driven by the mindset which states if it works, don't replace it, this is where vulnerabilities are found. Legacy protocols such as Modbus, Distributed Network Protocol (DNP3), and Profibus are designed for a technical ecosystem of trust. They lack the basics of cyber security such as authentication and encryption (Morris & Gao, 2013). A Modbus TCP packet is comprised of a transaction identifier, protocol identifier, length field, unit identifier, function code, and the data itself. What is concerning is that there is no field for a password or any sort of user authentication. If a bad actor can access the network segment at Level 1 where the packet transactions take place, they can inject a "Function Code 5" (Write Single Coil) command to control the valves or turn pumps off in a system. The PLC obeys these orders with no questions. The machines have no way to distinguish between a legitimate command that comes from a Human-Machine Interface (HMI) and a command from a hacker's laptop in St. Petersburg. The entire system of global infrastructure lacks any sort of "Origin Integrity" (Stouffer et al., 2015).

2.2 PLC Logic and Firmware Manipulation

At the center of every industrial process is the Programmable Logic Controller (PLC). These computers execute something known as "Ladder Logic," a symbolic programming language meant to represent electrical relay circuits. Advanced cyber warfare attacks like Stuxnet do not only send bad commands, but they can also infect the PLC itself. This strategy of "Logic Injection" attacks involve an attacker rewriting the bytecode that is compiled and ran on the PLC's microprocessor (Langner, 2011). Firmware Modification is a more insidious technique. Firmware is low level code that interfaces with the PLC's hardware by connecting to its operating logic. By overwriting firmware, a "Rootkit" can be created at the hardware level. This allows the PLC to be issued malicious commands (spinning a turbine too fast) while reporting a normal status to the SCADA monitoring screens. This attack is occurring on the sensor itself. The operator can be looking at a screen that reports 0 RPM while the machine is disintegrating at 5,000 RPM. This is Epistemological Warfare, which involves manipulating the operator's perception of reality (Langner, 2011).

2.3 Safety Instrumented Systems (SIS)

Critical infrastructure has a last line of defense known as Safety Instrumented Systems (SIS). SISs are autonomous control loops, they are physically separated from the main process control, designed for monitoring critical thresholds like pressure and temperature, if those reach alarming levels it initiates a emergency shutdown (ESD). The SIS is supposed to operate on a closed loop within control systems. Modern "Smart" SIS controllers like the Schneider Electric Triconex controllers like remote diagnostics. The Triton/Trisis malware specifically targeted the firmware of the safety controllers (Sobczak, 2019). The malware interacted with the TriStation protocol which is a UDP-based protocol that configures controllers. Reverse engineering checksum algorithms and command codes of the TriStation protocol enabled the attackers to inject a Remote Access Trojan (RAT) into the memory of the safety controller. This malware had a conditional logic bomb, if the main attack fails, it disabled the SIS. Disabling the SIS prevents the failsafe from activating, allowing the plant to explode.

3. Ethical Issues

Weaponizing ICS presents a challenge to Just War Theory by creating new categories of moral responsibility and liability for state actors and engineers.

3.1 Erosion of Distinction in Multi-use Grids

The Principle of Distinction found in International Humanitarian Law (IHL) states that belligerents must distinguish between military objectives and civilian activities (Schmitt, 2017). The Principle of Distinction is impossible to uphold if a nation state engages in cyberwarfare against critical infrastructure. A powerplant is considered "Dual-Use" due to a single turbine that can power a local hospital and a military radar at the same time. Electrons do not carry designation tags, therefore, a cyber attack cutting power with the intention of degrading military capabilities will also result in civilian suffering. This shifts the burden of Proportionality onto the attackers. The Foreseeable Harm of cutting off power in a region can cause traffic accidents due to shut off lights, food spoiling causing economic loss, and the failure of medical equipment which could result in causing death. In traditional kinetic warfare, a kinetic strike upon a specific electrical substation knowing it feeds a base can also affect a civilian population. A cyber-attack upon power grids could cause a "Cascading Failure" which can cause blackouts that reach across entire regions. The effects of attacks on ICS's are inherently indiscriminate, which in turn makes them unethical.

3.2 Sociotechnical Imperative and Negligence

Anticipatory Ethics would place upon the operators and designers of the Critical Infrastructure a large moral responsibility. The Sociotechnical Imperative states that systems must be designed for the environments they operate in instead of idealized ones (Miller et al., 2011). ICS vendors prioritized "Availability" over "Confidentiality" thinking their systems would never connect to the internet. If you connected a Windows-XP based HMI to the internet in the modern era through TeamViewer, seen in the Oldsmar hack (CISA, 2021), it would count as Criminal Negligence. Failing to segment networks or implement Unidirectional Gateways is an ethical failure of failing to foresee the misuses of technology (Brey, 2012). The engineers and administrators can be complicit in the attacks that target critical life-safety systems by deploying an inherently faulty design. The state bears moral responsibility for stockpiling "Zero-Day" exploits meant for offensive use targeting ICS systems, which prioritizes offensive military operations over the safety of its own populus.

4. Case Studies

The following case studies illuminate the evolution of ICS warfare which causes indiscriminate destruction, and which highlight specific technical mechanisms used to bridge the gap between code that exists only in digital space and the physical realities that are the targets of these attacks.

4.1 Stuxnet (2010)

Stuxnet used four Zero-Day exploits to spread through Windows networks until it found the specific workstations that ran Siemens Step7 software. Once Stuxnet was inside the software it replaced a DLL (s7otbxdx.dll) with its own harmful version**.** A DLL is a dynamic link library, a shared code file that is used by Windows systems (Microsoft, 2021). This allowed the malware to replace incoming commands sent by an engineer to the PLC. Malicious Statement List (STL) was injected into the Siemens S7-315 PLCs which controlled the centrifuges. The logic changed the output frequency of the variable frequency drives (VFDs) from 1,064 Hz to 1,410 Hz which caused the rotors to spin at supersonic speeds and shatter, then dropped to 2 Hz, and then moved back up again (Zetter, 2014). It recorded 21 seconds of normal sensor data and looped it to play on the monitoring screens (Level 2 SCADA), making sure the attack went under the noses of the operators (Karnouskos, 2011). Stuxnet not only met the threshold of an "Armed Attack" due to its physical destruction, but it was also arguably just under Utilitarian outcomes-based ethics as it delayed Iranian nuclear proliferation. It demonstrated the ability for precise distinction, targeting only the PLCs with specific Profibus network cards.

4.2 Ukraine Power Grid (2015 & 2016)

The attack on Ukraine's power grid (Kyivoblenergo, Prykarpattyaoblenergo) was conducted by the Russian "Sandworm" group (Greenberg, 2019). This attack represents a shift from a precision cyber-attack to a brute-force attack. The attackers gained entry to the system through spear-phishing with the "BlackEnergy 3" malware. They targeted the OT networks and stole VPN credentials. On December 23, 2015, they accessed SCADA Operator Workstations. A synchronized attack where they virtually moved the mouse on operator screens, opening the circuit breakers at 30 substations, effectively cutting power to 230,00 customers (Lee et al., 2016).

Normally someone can simply turn the braker back on, but the Sandworm group inserted KillDisk wiper malware to overwrite the Master Boot Record (MBR) of the operator workstations, making it so operators could no longer boot them. Not only did they effectively kill the MBR, but they also targeted Serial-to-Ethernet converters which are devices that translate IP packets to commands for legacy breakers and overwrote their firmware with garbage data "bricking" the hardware (Assante, 2016). This forced Ukrainian operators to drive to the substations and close the breakers manually in the middle of winter. This attack targeted civilian heating and light sources during winter, it was a direct violation of the prohibition on attacking objects that are required for the survival of a civilian population. Russia wanted to freeze the Ukrainian population into submission, which counts as psychological warfare.

4.3 Triton/Trisis (2017)

Perhaps the darkest chapter in ICS warfare was the targeting of the Petro Rabigh petrochemical plants in Saudi Arabia by targeting the Safety Instrumented System (SIS). Attackers reverse-engineered the TriStation protocol used on UDP port 1502 by targeting the Schneider Electric Triconex safety controller. A payload was injected into the firmware of the controllers. The malware logic monitored the system's memory for a "trigger" condition, that being if there was an emergency, the failsafe would be disabled (Sobczak, 2019). The only reason this attack failed was due to the malware containing a bug that made the Triconex controller crash into a "Failsafe" mode (stopping the plant) as opposed to running malicious code. Stuxnet represented sabotage, Ukrainian power grid shutdowns represented disruption, Triton had the intent to kill people. Disabling the safety system serves one purpose, catastrophe. First a catastrophic explosion, then toxic gas releases, finally, the toxic gas causes mass casualties in the surrounding area. This technically fits the definition of being a War Crime, targeting noncombatant workers and innocent civilian populations with kinetic explosion and chemical warfare.

4.4 Colonial Pipeline (2021) and JBS Meat Packing

These intrusions effectively attacked the supply chain. They demonstrate the vulnerabilities that arise when IT and OT connect, and the rise of Ransomware-as-a-Service (RaaS). In the case of the Colonial Pipeline, the professional DarkSide ransomware group managed to compromise the corporate IT network by stealing a VPN password for an inactive account, which was leaked on the dark web. The ransomware technically only infected the IT billing and administrative systems, Colonial proactively shut down the OT pipeline because they can not bill customers for the fuel (Sanger & Perlroth, 2021). This is best viewed with the Interdependency Vulnerability in mind; OT was stopped because of the finances relating to IT. Similarly, JBS was attacked by Revil by using data exfiltration and encryption to halt meat processing plants across three continents (Turton & Mehrotra, 2021). Economic entities are now proxy battlefields. Criminal groups (often state-tolerated or sponsored) are willing to hold national food and fuel supplies hostage for economic gain through extortion. This obscured the line that separated cybercrime from economic warfare. The state must now consider whether a ransomware attack is an existential threat that justifies a military response.

4.5 Oldsmar Water Plant (2021)

In February 2021 there was a near miss when with a cyber-attack a hacker accessed the SCADA system of a water treatment plant located in Oldsmar, Florida (Perlroth, 2021). The plant used the outdated Windows 7 operating system, which a year earlier stopped receiving security updates. The plant was running TeamViewer which is a commercial remote desktop software which was installed directly on the SCADA workstations, and running on outdated software (CISA, 2021). The attacker managed to take control of the mouse and went to the HMI screen which controls chemical dosing. They changed the dosing of Sodium Hydroxide (Lye) from 100 parts per million (ppm) to 11,100 ppm, which at these levels can cause severe chemical burns or even death. Luckily, an operator noticed the mouse moving on the screen and immediately reversed the changes. This demonstrates how low the barrier to entry is for ICS warfare. This was not a sophisticated attack by a state; it was likely a hacker using a shared password. Yet the kinetic effect being mass poisoning would have caused a catastrophe.

5. Anticipatory Ethics

Failing to prevent Critical Infrastructure from being weaponized is a severe ethical failure of foresight. By applying "Moral Responsibility for Computing Artifacts" framework developed by Miller et al. (2011), specific lapses in engineering and administrative decisions can be identified.

5.1 Rule 4: The Sociotechnical Imperative

The Sociotechnical Imperative states, "People who design... can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artifact is embedded" (Miller et

al., 2011). This rule was ignored in the Oldsmar water plant incident, The designers of the SCADA workstation created a system where it was possible to remotely mix lethal chemicals into water. Their system prioritized remote convenience and budget constraints over safety. The installation of TeamViewer, a remote desktop tool meant for consumers to use which was placed directly onto machines and which then controlled life-safety actuators is an immense oversight. Ease of access was placed higher than the people's safety. They did not anticipate that within a sociotechnical environment filled with credential stuffing and shared passwords, that their tool would be compromised. A responsible implementation of this rule would require an" Out of band" access method or a hardware-based VPN with multi factor authentication. They failed to recognize the weakest link in the security chain is the user. This decision allowed a random hacker to attain admin level control of chemical dosing within the system.

5.2 Rule 2: The Foreseeability of Effect

Rule 1 states "The people who design... are morally responsible for that artifact, and for the foreseeable effects of that artifact" (Miller et al., 2011). This can be applied directly to the state actors who hoard Zero-Day vulnerabilities for ICS attacks. The creators of the "EternalBlue" exploit that was used in NotPetya and the Triton framework saw these tools designed to interface with industrial protocols (Greenberg, 2019). It was foreseeable that once these weapons were deployed, they could be reverse-engineered or leaked which would allow for them to be used by criminal gangs and hostile state actors. The foreseeable effect of developing malware that targets Schneider Electric safety controllers is eventually a negative outcome since all safety controllers become vulnerable to the same exploit (Brey, 2012). The state decided to prioritize offensive capabilities for destructive purposes over the defensive duty to update the safety of its own infrastructure, violating the social contract between a nation and its citizens. When a government finds flaws in a SIS, they are ethically obligated to disclose these shortcomings and patch them.

5.3 The Post-Deployment Mandate

Rule 2 declares "Responsibility includes being answerable for the behaviors of the artifact... after deployment" (Miller et al., 2011). ICS's have lifecycles measured by decades. The "BlackEnergy" malware that was used in Ukraine to exploit legacy protocols was developed over 20 years ago. While those developing the protocols 20 years ago may not have been able to foresee these consequences, modern operators should have, they bear the post-deployment responsibility where one must maintain their systems in a hostile environment. The system being secure when you bought it is not valid under Anticipatory Ethics. Critical Infrastructure operators have a continuous moral duty to examine their systems to harden them against emerging threats. The Colonial Pipeline operators failed to update their primitive VPN profiles, outdated IT software, and this is a failure to uphold the Post-Deployment Mandate. They allowed an account no longer in use to exist, which creates a vulnerability that paralyzed the entire East Coast fuel supply. Liability does not end at the time of installation; it persists until the artifact is no longer used.

6. Recommendations

To effectively halt easy unsophisticated cyber-attacks in ICS warfare, the reactive patching methodology must be replaced. Instead, a secure-by-design policy must be adopted that enforces security at the physical, network and device levels. This requires the re-engineering of the OT systems, using deterministic hardware constraints and physical validation to eliminate cyber threats.

6.1 Hardware-Enforced Unidirectional Security Gateways

The most resilient defense for critical network segments involves the replacement of software level firewalls with hardware-enforced Unidirectional Gateways, commonly called data diodes. Firewalls rely on software rules that can be bypassed with zero-day exploits or a misconfiguration, data diodes enforce security at OSI Layer 1 (Physical Layer) (Stouffer et al., 2015). The device contains a fiber-optic transmitter on the OT side and a receiver on the IT side, but they lack the physical hardware capabilities to transmit light in the opposite direction. This is a physics-based air gap that allows the monitoring of data to flow out of the enterprise network for analysis while rendering it impossible for any packet, command, or malware to enter the OT network. This breaks the TCP/IP protocol which requires a bi-directional three-way handshake (SYN, SYN-ACK, ACK). To overcome this the diode uses proxy servers. The source proxy on the OT side mimics a TCP endpoint terminating the connection and getting rid of the payload. It converts the payload into a one-way UDP stream or serial frame across the optical link. The IT side receives this stream and reconstructs a valid TCP session for the application. Implementing this at Level 3.5 DMZ makes sure that even if the entire corporate network is compromised by a ransomware, adversaries cannot send the packets that are needed to exploit PLCs on the plant floor.

6.2 Deep Packet Inspection (DPI) and Industrial Protocol Sanity Checking

Passive network monitoring is currently a simple flow analysis, which needs to be switched to a Deep Packet Inspection (DPI). Current intrusion detection systems (IDS) typically ignore ICS packet payloads, only scanning the IP headers. A comprehensive DPI solution must parse the structures of the Modbus, DNP3, and Ethernet/IP packets to validate the legitimacy of the commands (Stouffer et al., 2011). A Modbus TCP packet that contains "Function Code 5" (Write Single Coil) is a real, legitimate command structure, but if it targets a register that is associated with a safety mechanism important to turbine overspeed valves, it will represent a hostile act. WE recommend deploying Context-Aware firewalls that enforce Stateful Logic. These systems would maintain a digital model of a plant's operational state. If a plant is undergoing maintenance, a command to update firmware may be permitted, if the plant is running, the DPI engine would cryptographically sign and drop any packet that attempts to write to firmware or critical memory blocks.

6.3 Physics Based Anomaly Detection (P-BAD)

We recommend the installation of a Physics-Based Anomaly Detection (P-BAD) system as an additional validation layer for SCADA data. Current Intrusion Detection Systems (IDS) search for the signatures of known malware. Stuxnet worked by looking like a real driver update. Stuxnet can not defeat the laws of physics. A P-BAD system would use a mathematical model of the physical processes (utilizing differential equations that represent fluid dynamics, thermodynamics and electrical loads) (Giraldo et al., 2018). The system compares the incoming data from PLCs against the expected physical reality that would be derived from the model. As an example, if a PLC reports that a valve has been closed, the laws of physics would state the pressure in the pipe must then increase. If the valve is closed and pressure remains unchanged, the P-BAD system would flag it as a Discrepancy. This would indicate a sensor is broken or an attack like Stuxnet is spoofing the sensor values. By evaluating the physics involved in the situation and not just the code, zero-day attacks can be defeated even if they bypass all digital security layers.

6.4 Firmware Integrity and Root of Trust

Manufacturers and operators of IC and ICS systems must enforce a hardware-based Root of Trust (RoT) by using Trusted Platform Modules (TPM). The Triton attack was successful due to Triconex accepting a firmware update without a comprehensive evaluation of the update. We recommend a Secure Boot protocol where PLC bootloaders calculate SHA-256 hashes of operating firmware, then compares it against a value burned into the chip's One-Time Programmable (OTP) memory during its manufacturing (Stouffer et al., 2015). If the hashes do not match, this indicates a rootkit or some sort of modification, the processor must refuse to boot and trip a Safe State relay at the physical level. We advocate for operators to regularly dump the memory of their PLC's and compare the Ladder Logic bytecode against a perfect image stored on a closed system. A change in the bytecode would indicate a Logic Injection attack. This level of forensic analysis is the only way to detect a resident malware that does not leave a file but lives inside the controller's volatile memory.

6.5 Out-of-Band Analog Safety Loops

Recognizing the fragility of all software, we recommend the re-introduction of Out-of-Band analog safety loops for life-safety functions. The Sociotechnical Imperative reminds us that efficiency should not come at the cost of survivability. In the Oldsmar water plant attack, digital controls allowed the lye levels to be increased to possibly lethal amounts. Digital safety limits in software can be overwritten by attackers. A physical defense or analog limit switch that operates on a closed system would prevent this. A centrifugal controller on a turbine that cuts the steam supply if RPMs exceed a certain physical threshold, or a chemical feed pump that is physically incapable through the limitation of the size of its impeller disallowing pumping more than a safe maximum flow rate would solve this problem (CISA, 2021). Designing physical constraints for machines leaving them with no choice but to operate at safe limits ensures that if the adversary attains system level privileges over digital controls, kinetic consequences are physically impossible to achieve, preventing catastrophe.

7. Future Work

Continuing research must address the integration of Artificial Intelligence into ICS. We anticipate the existence of Self-Healing Grids where AI monitors the physics of the plant and isolates compromised segments instantly (Active Defense). However, this introduces the risk of "Adversarial AI," where malware is designed to trick defense AI into shutting down the grid by intentionally feeding it faulty data. Researching "Physics-Informed Machine Learning," which allows for commands to be validated against the laws of thermodynamics before execution is the next frontier of ICS security.

8. Conclusion

The mixing of IT and OT has militarized the infrastructure civilians use in their daily life. The case studies detailed in this paper demonstrate that the "Air Gap" is dead. These are not just technical glitches; they are acts of aggression that weaponize critical infrastructure against us. Looking at it from a technical perspective, relying on insecure design protocols like Modbus and a lack of firmware integrity checks in PLC, allow for a wide array of attacks. Ethically, targeting multi-purpose infrastructure is a violation of the principles of Distinction and Proportionality in Just War Theory. Cyber conflict now operating in a "Gray Zone" allows states to attack heating, water, and the fuel supplies of adversaries with impunity that would not be tolerated in traditional warfare. We conclude that the security of Critical Infrastructure is a moral imperative. The Sociotechnical Imperative demands systems designed to prioritize safety over the convenience of remote administration. Until the protection of SCADA systems are treated with the same seriousness as the protection of a nuclear missile silo, we remain at risk of a catastrophe designed by ourselves.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.1 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing.

References

Assante, M. J. (2016). *Confirmation of a coordinated attack on the Ukrainian power grid*. SANS Industrial Control Systems Security Blog.

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. https://www.acm.org/code-of-ethics

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Cybersecurity and Infrastructure Security Agency (CISA). (2021). *Compromise of U.S. Water Treatment Facility*. Alert (AA21-042A). https://www.cisa.gov/uscert/ncas/alerts/aa21-042a

Giraldo, J., Urbina, D., Cardenas, A., Valente, J., Faisal, M., Ruths, J., Tippenhauer, N. O., Sandberg, H., & Candell, R. (2018). A Survey of Physics-Based Attack Detection in Cyber-Physical Systems. *ACM Computing Surveys*, 51(4), 1–36.

Greenberg, A. (2019). *Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers*. Doubleday.

Karnouskos, S. (2011). Stuxnet worm impact on industrial cyber-physical system security. *IECON 2011 - 37th Annual Conference of the IEEE Industrial Electronics Society*, 4490–4494.

Langner, R. (2011). Stuxnet: Dissecting a cyberwarfare weapon. *IEEE Security & Privacy*, 9(3), 49–51.

Lee, R. M., Assante, M. J., & Conway, T. (2016). *Analysis of the Cyber Attack on the Ukrainian Power Grid*. SANS Institute.

Microsoft. (2021, September 21). *What is a DLL?* Microsoft Learn. https://learn.microsoft.com/en-us/troubleshoot/windows-client/setup-upgrade-and-drivers/dynamic-link-library

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

Morris, T., & Gao, W. (2013). Industrial control system traffic data sets for intrusion detection research. *International Conference on Critical Infrastructure Protection*, 65–78. Springer.

Perlroth, N. (2021). *This Is How They Tell Me the World Ends: The Cyberweapons Arms Race*. Bloomsbury Publishing.

Rid, T. (2012). Cyber War Will Not Take Place. *Journal of Strategic Studies*, 35(1), 5–32.

Sanger, D. E., & Perlroth, N. (2021, May 8). Cyberattack Forces a Shutdown of a Top U.S. Pipeline. *The New York Times*.

Schmitt, M. N. (Ed.). (2017). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge University Press.

Sobczak, B. (2019). *The Inside Story of the World's Most Dangerous Cyberattack*. E&E News. https://www.eenews.net/articles/the-inside-story-of-the-worlds-most-dangerous-cyberattack/

Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M., & Hahn, A. (2015). *Guide to Industrial Control Systems (ICS) Security*. NIST. https://csrc.nist.gov/publications/detail/sp/800-82/rev-2/final

Tavani, H.T. (2009). *Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology*. Wiley.

Turton, W., & Mehrotra, K. (2021, June 2). JBS Cyberattack: Hackers Demand Ransom in Bitcoin. *Bloomberg*.

Williams, T. J. (1992). *The Purdue Enterprise Reference Architecture: A Technical Guide for CIM Planning and Implementation*. Instrument Society of America.

Zetter, K. (2014). *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*. Crown.

Smart Phones and Current Developments in Cyberwarfare: An Ethical and Anticipatory Ethical Analysis

Richard Wilson[1,2] and Noah Donnelly[2]
[1]Department of Philosophy, Towson University, Baltimore, Maryland, USA
[2]Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA

wilson@towson.edu
ndonnel1@students.towson.edu

Abstract: Smartphones in cyber warfare raise serious ethical concerns due to a number of factors including obscuring the line between civilian and military technology, how they expose non-combatants to harm, and how they lack clear international regulation. The central ethical issues related to the use of smart phones in cyber warfare include: (1) Civilian vs. Combatant Distinction. Smartphones are primarily civilian devices, yet they can be weaponized allowing civilians to engage in cyber-attacks. This development undermines the principle of distinguishing between combatants and non-combatants, a cornerstone of international humanitarian law. (2) Collateral Damage. Malware or cyber operations launched via smartphones can unintentionally spread to civilian networks, hospitals, or financial systems. Unlike traditional weapons, cyber tools are hard to contain, making unintended harm more likely. (3) Privacy Violations. Smartphones store vast amounts of personal data. Using them in cyber warfare risks mass surveillance, identity theft, and exploitation of private information, raising ethical questions about consent and proportionality. (4) Accountability and Attribution. Cyber-attacks via smartphones are difficult to trace. This creates ambiguity about responsibility, making it harder to hold aggressors accountable under international law. (5) Escalation of Risks. Since smartphones are ubiquitous, their use in cyber warfare lowers the threshold for causing a conflict. Everyday devices could become tools of state-sponsored attacks, increasing the risk of escalation into broader wars. (6) Lack of Regulation. Unlike conventional warfare, cyber warfare has no equivalent of the Geneva or Hague Conventions. The absence of agreed-upon rules leaves smartphone-based attacks in a legal and ethical gray zone. This analysis will identify the ethical and anticipated ethical issues with the use of Smart Phones in Cyberwarfare and the ethical and anticipated ethical issues with identifying smart phones as an important factor in cyber warfare.

Keywords: Cyber warfare, Smartphones, Stakeholders, Pegasus spyware, Civilian-Combatant distinction, Anticipatory ethics, Geolocation, Kinetic targeting

1. Introduction

Technology is moving towards integrating computing directly into human biology. We see this with new Brain Computer Interfaces (BCIs) like Neuralink. However, before addressing the ethics of a chip in the brain, we must first examine the ethics of the "supercomputer in the pocket" the smartphone. This paper argues that the smartphone ecosystem serves as a critical "ethical pilot program" for merging civilian life and military targeting. This pilot program is currently failing. The normalization of data vulnerability and the weaponization of consumer electronics have set a precedent. Civilian users are now treated as military sensors. By analyzing the current ethical failures of smartphones in cyberwarfare, we aim to establish a framework. This framework can prevent these same vulnerabilities from being hardwired into the human mind in the near future.

The development of Global Positioning Systems (GPS), high-resolution cameras, microphones, and constant internet connectivity have turned every smartphone into a possible node in military Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) networks (Singer & Brooking, 2018). The multi-use nature of these technologies creates a complicated ethical landscape. When a civilian uses a smartphone to film troop movements and uploads them to a social media platform, have they now become a combatant? When a state actor infects the smartphones of dissidents with professionally made spyware like Pegasus, does that constitute an act of war against another nation?

The importance and need for this analysis is shown by recent geopolitical events. In the war in Ukraine, smartphones have been used by civilians to report artillery targets, essentially crowdsourcing the “Kill Chain.” In the middle east, espionage software developed by private corporations have been used to assassinate journalists. In military bases, fitness tracking apps on smartphones have inadvertently revealed the locations of special forces outposts. These incidents demonstrate that smartphones are no longer just passive devices, they now allow their owners to become active participants in global conflicts. This analysis investigates the ethical and technical implications of this shift through four case studies: the Pegasus spyware developed by the NSO Group, Ukraine “Diia” app, the Strava Leaks, and the hacking of Jeff Bezos. This paper analyzes these incidents using the lens of Anticipatory Ethics and the ACM Code of Ethics. This paper argues that the designers of mobile

operating systems have failed to uphold the Sociotechnical Imperative by prioritizing convenience over security, jeopardizing the human user in a militarized world (Miller et al., 2011).

2. Technical Issues

To understand why the smartphone plays a central role in modern cyber warfare, one needs to examine the technical architecture within it which makes it both vulnerable and incredibly valuable to state actors for the purposes of cyber warfare. The smartphone is technically distinct from a laptop due to its intimacy, sensors, and connectivity.

The first significant vulnerability is **Sensor Fusion**. Modern smartphones are equipped with an array of sensors including accelerometers, gyroscopes, magnetometers, barometers, GPS receivers, and microphones. In a benign world, these allow for navigation, fitness tracking, and entertaining apps. In the context of warfare this is known as Sensor Fusion. A compromised smartphone can provide state actors with real-time, three dimensional, constantly updating images of the target's environment. Malware can activate the microphone to record classified conversations (SIGINT) or use the camera to gather visual intelligence (IMINT) unbeknownst to the user (Zuboff, 2019). The technical issue is these sensors are normally privileges at the hardware level, meaning that if a user disables "Location Servers" in the software, for security reasons, smartphone users do not have the kernel level permissions required to turn these off, meaning the baseband processor may still be communicating with cell towers, allowing for triangulation which reveals the location of the cell phone.

The second and most sophisticated form of smartphone warfare involves Zero-Click Exploits. Traditional malware requires user interaction, such as when a victim falls for bait in phishing. However, weapons like the NSO Group's *Pegasus* utilize vulnerabilities in the way smartphones process incoming data packets, such as an iMessage or a WhatsApp call, all of which happens before the user even receives a notification. This mode of attack targets the Baseband Processor, the chip that manages radio functions. Since the baseband runs a proprietary, closed-source real-time operating system (RTOS) with access to memory, exploiting this successfully would give the attacker the equivalent of System privileges on Windows (the most privileged account, not intended for human use). This technical reality renders traditional user and application-level defense mechanisms obsolete, there is no link to avoid clicking, thus no defense against the weapon.

A secondary technical threat is the physical supply chain and Hardware Implants. Smartphones manufactured in adversarial nations may contain hardware backdoors or firmware modifications designed to exfiltrate data to the opposing nations state intelligence services. This "interdiction" technique allows for the pre-positioning of cyber weapons within the civilian population of a rival nation. Since the global supply chain for smart phones is so complicated, components being sourced from dozens of countries, this in turn makes verifying the integrity of the hardware almost impossible. This creates a situation where the hardware acts as a hostile agent.

3. Ethical Issues

The weaponization of smartphones fundamentally changes the nature of warfare as defined in Just War Theory, more specifically, the principles of Jus in Bello (conduct in war). In Just War theory the main ethical framework for assessing these weapons is the conflict between military necessities and humanitarian protection.

The cornerstone of International Humanitarian Law (IHL), codified in the Geneva Conventions is known as the Principle of Distinction. Belligerents need to distinguish between combatants and non-combatants, which refers to valid and invalid targets respectively. Smartphones start to blur this line. If a civilian uses their smartphone to participate in a cyber-attack or to provide targeting data to a Telegram bot, they could potentially lose their protected status and become a legally legitimate target (Dawson & Innes, 2019). Unlike a soldier in uniform, a "digital combatant" may look exactly like a civilian. This creates an ethical crisis: if everyone has a smartphone, and every smartphone can be a weapon, is every civilian now a potential target? This logic leads to Indiscriminate Warfare, violating the core principles of humanitarian ethics as applied to warfare.

The Principle of Proportionality states that acts of warfare are "expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated" (International Committee of the Red Cross [ICRC], n.d.). Smartphone-based cyber weapons are incredibly difficult to contain. A "worm" designed to disable the phones of enemy officers that spreads by Bluetooth or Wi-Fi can infect the phones of hospital staff, aid workers, and children (Singer & Brooking, 2018). This "digital contagion" causes Collateral Damage not only to data, but to critical services that rely on mobile communications. Under Rights Ethics, this does not respect the right of security and safety afforded to civilians. Under ACM Principle 1.2 (Avoid Harm), the deployment of

uncontainable malware on civilian devices is an unethical renunciation of the developer's professional responsibility (Association for Computing Machinery, 2018).

Finally, the use of smartphones for mass surveillance violates Kantian Deontology. Kant argues that individuals must be treated as ends in themselves and not means. When a state transforms private devices of its citizens into a surveillance grid to root out insurgency or dissent, it treats its populus as a mere means and source of data. This destroys the concept of consent in regard to privacy (Nissenbaum, 2009). The user consented to carry a phone for communication, and not for the purpose of being a tracking beacon for a government. This is a breach of the "Digital Social Contract" and represents an immense ethical failure of both state and technology providers. This failure is worsened by "Normalization." The public has learned to accept total surveillance, seeing it as the "cost of doing business" in the digital age. This complacency creates a positive feedback loop. Civilians who accept vulnerability in peacetime become defenseless targets in wartime.

4. Case Studies

The theoretical risk of smartphone warfare has already materialized in many high-profile conflicts and worldwide espionage operations.

The most defining example of the weaponization of smartphones in the modern era is the "Pegasus" suite developed by the Israeli cyber-intelligence firm called the NSO group. Unlike regular malware which relies on "social engineering" (tricking the user into clicking on a malicious link), Pegasus introduced a new era of "Zero-Click" exploits. The technical sophistication of Pegasus represents a massive leap in offensive cyber warfare capabilities. In 2021, researchers at the Citizen Lab discovered the "FORCEDENTRY" exploit that targeted Apple's iMessage architecture. The exploit involved the sending of a malicious PDF file that was disguised as a GIF image to the intended victim's device. Inside the file was a stream of data encoded in the JBIG2 compression. The NSO Group engineers found that the JBIG2 vulnerability in Apple's "CoreGraphics" library was so extensive that they could build a primitive, emulated computer inside the memory of an iPhone's image processor. This let them execute malicious code by the phone trying to render a thumbnail which requires no interaction from the user. Once installed, Pegasus grants the attackers root access privileges. It can steal passwords, read encrypted messages on WhatsApp and Signal, track a user's location history, and activate the microphone and camera remotely. The unethical failures were revealed by the "Pegasus Project," when investigative journalists revealed that the software was not just used against terrorists, but against heads of state, human rights activists, and journalists. A prominent victim was named Hanan Elatr, who was the wife of murdered journalist Jamal Khashoggi. A forensic analysis showed that her Android phone was targeted by UAE officials using the NSO software months before her husband was murdered. This demonstrates how smartphones can be used by totalitarian regimes to extend their repression beyond their own borders.

Unknown to the world, the 2022 Crowdsourced Kill Chain of the "Diia" app and "GIS Arta" would spark what has be referred to as the first "Smartphone War." The Ukrainian government switched its pre-existing e-governance app, "Diaa" (Action), into a national defense weapon. They launched a feature known as "eVorog" (e-Enemy), allowing for any citizen with a smartphone to report the geolocation of Russian military assets. To verify authenticity, users had to sign in with their digital ID. This data was subsequently fed into "GIS Arta." GIS Arta collects civilian reports, drone feeds, and NATO intelligence to feed locations of targets to the nearest artillery battery, thus reducing the time from identifying a target to firing from 20 minutes to less than one minute. While it is an innovative strategy, this creates an enormous Anticipatory Ethical crisis related to determining the nonmilitary or military status of civilians. According to the Geneva Conventions, a civilian loses their protected status "for such time as they take a direct part in hostilities." By using smartphones to guide artillery fire, a Ukrainian grandmother can now technically become a combatant. This blurs the "Principle of Distinction" so far that it becomes nearly irrelevant within a modern battlefield. This also has the unintended consequence of inviting Russia to treat the entire cellular network as a military command-and-control system, which would justify broad kinetic attacks on cell towers and internet infrastructure, potentially going so far as to target of individuals holding smartphones in or around combat zones. This case study shows the transition of the smartphone from a device for communication to becoming a weapon in the "Kill Chain" in military operations (Claverie & du Cluzel, 2022). This creates a "Gray Zone" related to combatant status that current International Laws are not equipped to address.

In 2018, the fitness tracking application known as **Strava** released a "Global Heatmap" that visualized 13 trillion GPS data points collected from its users. While its intention was to show popular jogging routes, the map inadvertently unmasked secret military operations. Security analysists noticed in remote areas of Syria, Niger,

and Afghanistan, regions that should be dark, that there were bright data points with perfectly defined loops and lines. These were the jogging routes of U.S. and allied soldiers running the perimeter of what were supposed to be secret Forward Operating Bases (FOBs). This shows the dangers of "Passive Leakage" and "Pattern-of-life" analysis related to smart phones. The soldiers were not victims of being hacked, they signed a End User License Agreement (EULA) agreeing to the upload of their data. As a result, the collection of this data did not simply reveal where the bases were, but the operational tempo and supply routes of the forces inside these bases as well. By analyzing the timestamps of the runs, an adversary could gain an insight into patrol schedules and when shift changes occurred. This was a failure of the Sociotechnical Imperative by both the app developers and the commanding officers. Strava did not anticipate that when users of the tracking device set their participation to "public" by default they could compromise national security. In this case military commanders did not anticipate that a consumer device could be used as a tracking beacon.

The hacking of Amazon CEO Jeff Bezos in 2018 showed the world that the smartphone is a mode for the execution of high-level coercive diplomacy. According to a forensic analysis by FTI Consulting, Bezos' iPhone X was breached after receiving a WhatsApp video file sent from then personal account of the Saudi Crown Prince Mohammad bin Salman (Kirchgaessner, 2020). In a matter of hours of receiving the video, the device began sending out massive amounts of data, increasing from a daily average of 430KB to 4.6GB. This attack was not motivated by financial theft, but instead was aimed at gaining political leverage. At the time, *The Washington Post* (owned by Bezos) was aggressively reporting the murder of Jamal Khashoggi. The hack was meant to gather compromising material to blackmail Bezos, in order to have him give hush orders to his reporting staff. This incident single handedly destroyed the myth of "High-Value Target" immunity. At the time Bezos was the wealthiest man on earth, having access to the best cybersecurity resources, but was still vulnerable to a state sponsored "Zero-Day" exploit. It redefined smartphones as a weapon for Economic Warfare and political coercion.

5. Anticipatory Ethics

Applying the rules designed for assessing "Moral Responsibility for Computing Artifacts" by Miller et al. (2011) to the smartphone ecosystem can reveal significant ethical lapses in judgement.

The Sociotechnical Imperative states that "People who design... can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artifact is embedded" (Miller et al., 2011). Apple and Google designed iOS and Android for a sociotechnical system defined by "convenience" and "connectivity." By forging an ecosystem that prioritizes data collection for advertising and background connectivity for updates, they inadvertently built an architecture that is at its root vulnerable to surveillance. A responsible design for a "Warfare-Ready" world would include hardware kill-switches for microphones and cameras (Brey, 2012). Failing to include these features in their operating systems is a failure to anticipate the adversarial environment in which Apple and Google are embedding their your products.

Rule 1, The Foreseeability of Effect argues that "The people who design... are morally responsible for that artifact, and for the foreseeable effects of that artifact" (Miller et al., 2011). It was foreseeable that GPS tracking data from apps like Strava could be used for intelligence gathering. It was foreseeable that "Zero-Click" vulnerabilities in messaging apps like WhatsApp could be weaponized by state actors. The developers bear moral responsibility for prioritizing the release of new features at high speeds to attract consumers over security auditing to protect the customers who are buying their products. The failure to patch these vulnerabilities before they were weaponized and sold to entities like the NSO Group, demonstrates how developers neglected the foreseeable harm caused by their product. The "move fast and break things" philosophy of companies such as Facebook is indefensible when ethically scrutinized when the thing being broken is the product sold to customers whose physical safety of threatened.

Rule 2, What is called the Post-Deployment Mandate states that "Responsibility includes being answerable for the behaviors of the artifact... after deployment" (Miller et al., 2011). NSO Group likes to argue that they sell the software but are in no way responsible for its use. Anticipatory Ethics voids this defense. For a socially responsible creator of a cyber weapon this responsibility should also extends to post-deployment use. If the smart phone becomes a weapon and is used to target civilians, the creator of the smart phone shares in the user's culpability. Similarly, app stores have a post-deployment mandate to audit apps that serve as fronts for state surveillance, such as TikTok. The failure to monitor the ecosystem for weaponized applications is a clear neglect of the ethical responsibility one bears once the product has left the factory (Collingridge, 1980).

6. Recommendations

To begin, we must acknowledge the "Realist" critique. These security measures remain unimplemented because they contradict the current business models of surveillance capitalism and the strategic interests of state intelligence agencies. A hardware "kill switch" reduces the data collection capabilities advertisers rely on, and end-to-end compartmentalization hinders the ability of state actors to monitor threats. We contend that these recommendations, though commercially inconvenient, are ethically mandatory. As technology moves from "pocket" to "implant," the cost of maintaining this vulnerability gap will shift from financial loss to the loss of human cognitive autonomy.

The first recommendation is the implementation of a Hardware-Level “Kinetic Mode”. Manufacturers should implement a physical switch on devices, distinct from the software-based “Airplane Mode,” which can easily be bypassed by any attacker with deep systems knowledge. This switch would physically disconnect the circuit to the GPS, Microphone, and Baseband transmitter, only allowing locally stored maps and data features to function. This restores the Principle of Distinction by allowing a user to have a device that is verifiably “silent” and thus can not be targeted, avoiding collateral damage (Tavani, 2009). It moves the control of the device’s emissions from the software layer, which is hackable, to the hardware layer, which can provide a mechanical, unbreachable Killswitch.

The second recommendation is the adoption of a Zero-Trust Supply Chain. Defense contractors and Governments must be required to adopt a model where no component of a device is trusted by default. Smartphones used in sensitive contexts should be manufactured in secured, domestic facilities with firmware that can be audited comprehensively. This would uphold the ACM Principle 2.9 *Robust Security* and prevent the risk of foreign intelligence assets accessing domestic infrastructure. Without this, the hardware itself remains and must be treated as, a potential Trojan Horse.

The third recommendation is International Cyber-Arms Trade Regulation. The international community must treat “Cyber Weapons” such as Pegasus with the same scrutiny as kinetic weapons (ITAR). The sale of zero-click exploits to regimes that carry a poor human rights record must be banned from trade. Just War Theory and the goal of minimizing suffering align with such regulations. Limiting the spread of these weapons, could substantially reduce the likelihood of their use against non-combatants (Alabama Policy Institute, 2020).

The fourth recommendation is Operating Systems Compartmentalization. Mobile operating systems currently operate in the “Sandbox” model. We propose a shift to a strict Hypervisor-Based Virtualization model, similar to Qubes OS on desktops. In this architecture, the “Baseband” (radio), “Personal Space” (photos, chats, apps), and the “Work/Military Space” (sensitive data) would run in their own individual, completely isolated virtual machines (VMs). As smartphones currently stand, a zero click exploit in the iMessage parser (Personal Space) will increase its privileges to take over the entire kernel of a smart phone. In a virtualized environment, a malicious PDF exploiting the message parser would be trapped in the personal space with no access to the microphone, GPS, or other VMs. Furthermore, these virtual machines would not persist after a reboot, which would neutralize the lethality of weapons like Pegasus. This aligns with ACM Principle 2.9 *Designed for Robustly and Usably Secure Systems*. Assuming code will inevitably have bugs, this architecture minimizes the possible harm caused by these bugs, which fulfills the Foreseeability of Effect rule by designing for resilience rather than designing for perfection.

The fifth recommendation is Civic Digital defense Training. Since civilians are now participants in the “Kill Chain” (through apps like Diia), states have a moral obligation to provide Cognitive and Technical Defense Training. This would include public education campaigns related to a person’s own signal monitoring (e.g., why you shouldn’t use fitness trackers on base), spotting social engineering, and the risks of metadata. The “Human Factor” is and always will remain the weakest link. Even a perfectly secure architecture can be compromised if the user volunteers their location data to Strava. Training would reduce the “attack surface” by educating the population, thus making them harder to trick. If the state relies on the population for intelligence, it owes them the tools to survive the retaliation caused by their actions. Under Rights Ethics, not training civilians while encouraging them to use their phones for war is a dangerous form of negligence which allows them to be exposed to foreseeable harm.

The sixth and final recommendation is Legal Protections for Security Researchers. We recommend the establishment of internation at protection laws for security researchers who discover vulnerabilities in digital infrastructure. Currently, researchers who find flaws (like in the Strava leak) often operate in legal grey zones, and fear prosecution under laws like the Computer Fraud and Abuse Acts (CFAA). “White Hat” hackers are the

white blood cells of the internet, they find vulnerabilities before adversarial state actors can turn them into weapons. Punishing security researchers could create a ripple effect that leaves vulnerabilities unpatched. This analysis is supported by ACM Principle 1.5 *Respect the work required to produce new ideas* and 1.2 *Avoid Harm*. Encouraging a responsible disclosure of vulnerability will ensure design flaws are fixed, protecting the population as a whole. Any legal framework should view researchers as "Digital First Responders" as opposed to criminals.

7. Future Work

As stated in the introduction, the ethical failures observed in the smartphone era serve as a warning for the upcoming era of Brain-Computer Interfaces (BCIs). Future research should examine the ethical implications of Brain-Computer Interfaces (BCIs) and Augmented Reality (AR) contact lenses. If the smartphone is integrated directly into the human brain, such as a product like Neuralink, a "hack" could be considered a cognitive assault. A "Zero-Click" exploit in this new context could theoretically write data to the visual cortex, causing hallucinations or altering an individual's perception of reality. Additionally, future devices will track biometric data to infer a person's psychological states, by analyzing things like heart rate and cortisol levels. This enables "Pre-Crime" targeting. Pre-Crime targeting would target a soldier or civilian not because of where they were or what they did, but because of how they feel. This would represent the ultimate convergence of biology and technology in cyber warfare.

8. Conclusion

The smartphone has inherently changed the shape of the battlefield. It has democratized the gathering of intelligence, allowing civilians to act as sensors for military operations, but in effect they have also democratized targeting, turning every civilian into a potential beacon for a missile strike. The blurring of the distinction between a "Civilian Device" and a "Military Sensory" is a defining ethical crisis of modern cyber warfare related to the use of smart phones.

Through the analysis of Pegasus, Ukraine, and Strava, we conclude that the existing state of smartphones and how they are used in war are incompatible with the principles of International Humanitarian Law. Relying solely on software switches and closed-source basebands creates an environment of "Security Nihilism," where no user can be sure of their privacy and protected status. By applying Anticipatory Ethics, we identify that the burden lies with the architects of the technology being used. We need to abandon the "Consumer-Grade" mindset when designing devices that are currently being used in military environments. The Sociotechnical Imperative demands that we design devices which protect the user not just from potential spam, but from military targeting. Failing to do this ensures that in future wars, the existence of non-combatants with protected status, may no longer exist.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.1 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing.

References

Alabama Policy Institute. (2020, November). *Understanding the Difference Between Positive and Negative Rights*. Alabama Policy Institute.

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. https://www.acm.org/code-of-ethics

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Claverie, B., & du Cluzel, F. (2022). Cognitive Warfare: The New Battlefield Exploiting Our Brains. *Polytechnique Insights*.

Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press.

Dawson, A., & Innes, M. (2019). How Russia's Internet Research Agency Built its Disinformation Campaign. *The Political Quarterly*, 90(2), 245–256.

Federal Register. (2025). *Preventing Access to U.S. Sensitive Personal Data and Government-Related Data by Countries of Concern* (89 FR 1636). U.S. Department of Justice.

Google Threat Intelligence Group. (2025). *Hello 0-Days, My Old Friend: A 2024 Zero-Day Exploitation Analysis*. Google Cloud. https://cloud.google.com/blog/topics/threat-intelligence/2024-zero-day-trends

Harvard Kennedy School. (2025). *Ukraine's Digital Transformation: Innovation for Resilience*. Center for International Development. https://www.hks.harvard.edu/centers/cid/voices/ukraines-digital-transformation-innovation-resilience

International Committee of the Red Cross. (n.d.). *Proportionality*. How does law protect in war? - Online casebook. https://casebook.icrc.org/a_to_z/glossary/proportionality

Kirchgaessner, S. (2020, January 21). *Amazon boss Jeff Bezos's phone 'hacked by Saudi crown prince'*. The Guardian. https://www.theguardian.com/technology/2020/jan/21/amazon-boss-jeff-bezoss-phone-hacked-by-saudi-crown-prince

Kirchgaessner, S. (2024, November 15). *NSO – not government clients – operates its spyware, legal documents reveal*. The Guardian. https://www.theguardian.com/technology/2024/nov/14/nso-pegasus-spyware-whatsapp

Knight First Amendment Institute. (2024, November). *Spyware Company NSO Group Faces Setbacks in Attempts to Avoid U.S. Lawsuits*. Columbia University. https://knightcolumbia.org/blog/spyware-company-nso-group-faces-setbacks-in-attempts-to-avoid-us-lawsuits

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

Miller, Maggie. (2025, May 6). *Israeli spyware giant NSO Group ordered to pay nearly $170M to WhatsApp for hacking accounts*. POLITICO.

Nestor, M.W. and Wilson, R.L. (2022). *Anticipatory ethics and the use of CRISPR in humans*. Springer.

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Proctor, J. (2022, August 6). Dutch man Aydin Coban found guilty of extortion in Amanda Todd case. *CBC News*.

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The Weaponization of Social Media*. Eamon Dolan/Houghton Mifflin Harcourt.

SOC Prime. (2025, November 5). *CVE-2025-48593: Critical Zero-Click Vulnerability in Android Enables Remote Code Execution*. SOC Prime. https://socprime.com/blog/cve-2025-48593-vulnerability-in-android/

System and kernel security. (n.d.). *System and kernel security*. Android Open Source Project.

Tavani, H.T. (2009). *Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology*. Wiley.

What Is a Rendered Image on iPhone. (n.d.). *What Is a Rendered Image/Video on iPhone*.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Russia, Weaponized Social Media and Cyber Warfare: Ethical and Anticipated Ethical Issues

Richard Wilson[1,2] and Noah Donnelly[2]
[1] Department of Philosophy, Towson University, Baltimore, Maryland, USA
[2] Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA

wilson@towson.edu
ndonnel1@students.towson.edu

Abstract: Russia's weaponization of social media refers to how Russian state-associated actors used and continue to use social media platforms such as Facebook, Twitter, Instagram, and YouTube to influence public opinion, spread disinformation, and destabilize societies, especially during elections and political crises. This strategy relies on multiple essential tactical components. The first tactical component is Disinformation Campaigns, where Russian operatives created fake accounts and groups posing as ordinary citizens to spread falsified stories on politically divisive topics (race, immigration, gun rights) to drive up political polarization. Secondly are Troll Farms like the Internet Research Agency (IRA), which employed thousands of people to post inflammatory content with the aim of manipulating online discourse. The third strategy is Bot Networks, utilizing automated accounts to amplify hashtags and create the illusion of widespread support for fringe narratives. The fourth strategy is Microtargeting, where malicious actors purchased ads exploiting emotional triggers and psychographic data to influence voter behavior. Finally, these campaigns Exploited Algorithms that prioritize engagement, making sure sensational content became viral. The goal of this weaponization is to destabilize trust in democratic institutions, amplify social division, and influence election outcomes to favor Russia's geopolitical interests. This analysis identifies the ethical and anticipated ethical issues with the Russian weaponization of social media as a form of Cyber Warfare. There is an interdisciplinary method employed in this analysis that draws upon distinctions taken from computer science, conceptual ethical analysis and case studies.

Keywords: Cyber warfare, Information operations, Disinformation, Internet research agency, Cognitive warfare, Anticipatory ethics

1. Introduction

In the contemporary realm of global conflict, the battlefield has expanded beyond the standard domains of land, sea, air and space to include the cyber domain and audience cognition. Whilst cyber warfare has historically focused on attacks on physical infrastructure or the theft of classified data, a new paradigm has emerged: the weaponization of social media to target psychological and sociological infrastructures of a nation. These phenomenon, best displayed by the operations of Russian state-sponsored actors, represent a massive shift in the nature of warfare. They move the objective of warfare from the destruction of physical assets to the disruption of the "OODA Loop"(Observe, Orient, Decide, Act) of an entire population (Claverie & du Cluzel, 2022). By invading the infosphere of a target nation, an adversary can manipulate public perception, erode trust in institutions, and incite civil unrest. This strategy is known as Cognitive Warfare, and it utilizes the architecture of the modern internet as a medium for the use of their weapons systems to produce mass influence (NATO ACT, 2021).

The significance of the weaponization of social media is highlighted by the combination of sophisticated psychological operations with advanced technological capabilities. In the past, propaganda was limited by the speed of broadcasted messages through legacy media and the ability of states to control information channels. Today the internet allows for micro-targeting of individuals based on psychological profiles which allows state actors to bypass the traditional media gatekeeping, delivering tailor fit news direct to users (Zuboff, 2019). This capability was demonstrated during the 2016 United States Presidential Election, where the Internet Research Agency (IRA), a Russian "troll farm," executed an extensive information operation designed to sow discord among citizens and influence the election (Almond et al., 2022). The 2016 election offers a foundational case study for these tactics. Russian strategy has since evolved significantly. The 2022-2025 invasion of Ukraine marked a shift from "covert disruption" to "overt information warfare." This newer approach uses deepfakes, AI-generated propaganda, and the weaponization of platforms such as Telegram and TikTok. This paper analyzes this evolution. It contrasts the manual "troll farms" of the past with the automated "cognitive warfare" of the present.

The clear distinction between "influence" and "warfare" in this context grows increasingly blurry. Traditional defenses of war require a kinetic act or a direct violation of a nation's sovereignty. However, when a foreign state decides to influence the public discourse of another nation with the intent to destabilize its institutions

and provoke internal conflict, it meets the strategic criteria of warfare (Dawson & Innes, 2019). The intent of social media warfare is to cause irreversible harm to the targeted state's ability to function. The "cyber weapons" being deployed are not boots on the ground but code, bots, and algorithms**.** The impact is the degradation of sovereignty and the ability to autonomously rule one's people, along with weakening the targeted nation on the geopolitical stage. This analysis investigates specific tactics employed by Russian state actors, including disinformation campaigns, troll farms, bot networks, and the exploitation of algorithms to accomplish their goals. It analyzes these actions through the lens of standard ethical principles, Anticipatory Ethics and the ACM Code of Ethics. It argues that the designers of social media platforms have failed to uphold the Sociotechnical Imperative by ignoring the possibility of the weaponization of their tools (Miller et al., 2011). Furthermore, it proposes that the defense against this form of warfare requires a move from reactive moderation to a proactively oriented cognitive cybersecurity.

2. Technical Issues

To understand how social media platforms have developed into weapons of cyber warfare, one needs to understand the technical architecture that facilitates this weaponization. The core technical issues lay in the design of the "Attention Economy" and the specific algorithms that govern user interaction on platforms including but not limited to: Facebook (now Meta), Twitter (now X), and YouTube. These platforms are not designed with civic integrity or national security in mind. Social media platforms are designed to maximize user engagement and length of engagement to maximize advertisement revenue (Zuboff, 2019). This commercial objective creates a set of technical vulnerabilities that state actors can exploit with devastating effects.

The primary mechanism of exploitation is the Recommendation Algorithm. These algorithms employ machine learning models, specifically Reinforcement Learning, to predict what keeps a user scrolling or clicking. Success is measured by "engagement," defined as likes, shares, comments, and engagement time. Research in behavioral psychology continuously shows that content on social media platforms which elicit high-arousal emotions, including fear, anger, and outrage tend to generate significantly higher user engagement than neutral or positive content (Bipartisan Policy Center, 2023). Consequently, the algorithms of profit driven social media platforms prioritize divisive, sensational, and inflammatory content. Russian operatives did not need to hack any code to spread their message, they simply needed to create content that triggered these algorithmic preferences in targeted audiences. By posting polarizing material, they ensured the platforms' automated systems proliferated their propaganda to millions of users with minimal economic expenditure. This phenomenon is known as "Algorithmic Radicalization," where the system pushes users towards increasingly extreme content to maintain engagement (National Institute of Justice, 2021).

A second crucial technical component of social media warfare is the capacity for Microtargeting and Psychometric Profiling. The business model of social media is built on "Surveillance Capitalism," where user data is harvested and used to create comprehensive profiles pertaining to individual behavior, preferences, and psychological traits (Zuboff, 2019). Advertisers then use this data to target specific demographics. Russian actors weaponized social media commercial tools by purchasing political advertisements targeting users who share specific vulnerabilities (Young & Kim, 2018). An example of this process would be identifying users who score high on "neuroticism" or "authoritarianism" in psychometric models and deploy ads to trigger aggression and paranoia in these audiences. This represents the industrialization of psychological warfare. In classic PSYOPS, leaflets are dropped on a population with a single, generalized message. In contrast, in social media warfare, a specific message is made for and directed towards specific individuals based on their user data which maximizes psychological impact while remaining unrecognized by a broader audience. This "Dark Post" (Singer & Brooking, 2018) architecture allows for reality to be fragmented where different sectors of the population are presented entirely unique narratives, making a general consensus on important topics nearly impossible to achieve in the overall population.

The third technical pillar is the deployment of Automated Bot Networks. A "bot" is a software application that runs automated tasks over the internet. In the context of cyber warfare, bots are employed to simulate human activity on social media. Russian actors deploy armies of bots to artificially boost specific narratives (Dawson & Innes, 2019). By programming thousands of accounts to "retweet" and "share" a specific piece of disinformation, they can manipulate the "Trending Topics" algorithms like those seen on Twitter. This creates a "Bandwagon Effect," where real human users see a trending topic and assume it represents popular opinion on breaking news (Singer & Brooking, 2018). This technique is known as "Astroturfing." Astroturfing creates an artificial illusion of widespread grassroots support for a fringe or completely fabricated idea. The technical failure here lies in the

inability or unwillingness of social media platforms to effectively distinguish between human and "bot" traffic. This creates an environment where automated narratives produced by Bots can drown out authentic discourse.

3. Ethical Issues

The weaponization of social media in the service of cyber warfare raises profound ethical questions that go beyond the technical details of the attacks. These ethical issues concern the inalienable rights of citizens, the moral obligations of corporations, and sovereignty in the digital age. This analysis applies ethical frameworks to the three primary stakeholder groups: the State Actors (Russia), the Platforms (Meta/Twitter), and the Citizens targeted through social media (victims).

From the perspective of International Relations and Political Ethics, the actions of Russian state actors violate the Principle of Non-Intervention and National Sovereignty. Sovereignty is the right of a nation to govern itself without external influence. The democratic process, specifically the election of a nation's leadership, is the most important expression of sovereignty. By covertly interfering in the 2016 U.S. election through information warfare, Russian actors violated the right of the American people to self-determination (Dawson & Innes, 2019). Rights Ethics defends the view that state citizens have a "Right to Mental Integrity" and a "Right to the Truth." Disinformation campaigns that deliberately employ social media to corrupt the information environment violate the epistemic rights of civilians (Tavani, 2009). Disinformation acts like a "virus" that exploits cognitive vulnerabilities, and which then deflects citizens from access to the accurate information required to make rational and informed political decisions. When state actors spread false news stories, such as claiming that a candidate is involved in criminal activity when thy aren't, they are not exercising free speech, they are engaging in a deception operation designed to subvert the autonomy of voters. This represents a violation of the Kantian duty related to honesty. Kantian Deontology argues that lying is inherently immoral because it treats the listener not as a rational agent, but a means to an end (Brey, 2012). Russian disinformation campaigns treated American voters as objects to be manipulated for geopolitical gain, which is a violation of their human dignity.

The ethical responsibility of the social media platforms is equally significant when evaluated through the lens of Virtue Ethics and the ACM Code of Ethics. Companies like Facebook (Meta) and Twitter (X) regularly claim to be neutral "town squares" that merely host content. This defense ignores their active role in curating and amplifying the content appearing on them using algorithms. The ACM Code of Ethics Principle 1.2 declares that computing professionals must "Avoid Harm" (Association for Computing Machinery, 2018). By designing algorithms that value engagement over veracity, and by allowing their platforms to be used for mass manipulation, these companies failed to avoid harm. The harm in this context is not only distress caused to individuals who were harassed by trolls, but systematic damage to the democratic ideal. The degradation of trust in institutions, increasing polarization, and incitements of violence are foreseeable harms that are enabled by the platforms' negligence of content moderation (Singer & Brooking, 2018).

Furthermore, the platforms violated ACM Principle 1.3, which necessitates the duty to "Be Honest and Trustworthy" (Association for Computing Machinery, 2018). For years, platforms have denied or minimized the extent of foreign interference being conducted on their platforms. They allowed bots to pose as humans, deceiving their users about the nature of interactions they have with Bots and the content spread by those Bots (Dawson & Innes, 2019). Virtue Ethics would ask, "what is the character traits of the designers of a social media system?" A system that covertly sells user vulnerabilities to foreign actors and prioritizes profit over the integrity of the democratic process lacks civic virtue. The concept of "Corporate Social Responsibility" (CSR) states that companies have an obligation to the societies within which they operate. By profiting from the capital spent on ads by Russian disinformation agents, these platforms directly profit from the destabilization of the nation's that host social media platforms (Quigley, 2017). This represents a failure of the social contract between the technological sector and the members of the public using social media platforms. These companies are placing shareholder interests, values and profit above the preservation of democratic principles.

4. Case Studies

The complex nature of "cognitive cyberwarfare" is best understood through the tactical implementations deployed during the 2016 U.S. election and subsequent geopolitical events. A series of case studies demonstrate the methodology under which the Internet Research Agency (IRA) and Russian Intelligence services (GRU) operate.

4.1 Disinformation Campaigns

The first prong of the Russian strategy was the industrial-scale production of disinformation. Russian operators made thousands of fake accounts, pages, and groups posing as American citizens or activist organizations. Groups with names like "Blacktivist," "Heart of Texas," and "United Muslims of America" were designed with the intention of infiltrating genuine political movements (Young & Kim, 2018). Once these pages had gathered authentic users, the operators would deploy "payload" content, that were fabricated or highly misleading stories designed to produce outrage. For example, during the 2016 election, Russian controlled pages spread false reports about voter fraud, criminal investigations into candidates, and fabricated quotes (Almond et al., 2022). One memorable operation involved Russian operatives creating two opposing protest events in Houston, Texas, on the same day. Another protest supported an Islamic center, while simultaneously protesting against the center. The goal was to increase political polarization by employing digital means related to an event that already involved a physical confrontation. Both events were organized remotely by Russian state actors in St. Petersburg. This tactic mirrored a Cold War strategy called "Active Measures," which was accelerated by the speed of the internet. The strategy relied on "Reflexive Control," (cite source) a concept in Russian military theory where the aim is to compel the opponent to make decisions. Decisions about how to vote or whether to engage in a protest are influenced by false premises and disinformation introduced by the attacker (Claverie & du Cluzel, 2022).

4.2 Troll Farms

The organizational heart of this warfare is the Internet Research Agency (IRA), a shadowy corporation based in St. Petersburg. The IRA operates as a "Troll Farm," employing over a thousand people to work 12-hour shifts creating content for social media. The facility was organized in a fashion similar to a professional marketing agency. The IRA had a graphics department, data analysis teams, and IT support (Dawson & Innes, 2019). The "trolls" job was to manage multiple fake personas across different time zones to simulate authentic American users. Their tactic was "comment warfare," where they would swarm the comment sections of news articles and Facebook posts to inject divisive rhetoric. They would regularly play both sides of an argument, having each side posting extremist viewpoints from both the left and the right to eliminate any middle ground, making accurate comprehension of messages impossible. This is essentially a "divide and conquer" strategy adapted to fit the digital age. The immense volume of content produced by the IRA created a "Firehose of Falsehood," which is a propaganda technique where the audience is overwhelmed by the high frequency of conflicting messages which can cause the audience to become cynical and disengage from politics entirely (Singer & Brooking, 2018). This achieves two strategic goals, 1st, the weaponizing of social media and voter suppression and the 2nd contributing to the erosion of faith in democratic discourse.

4.3 Bot Networks

While human trolls created messages, automated bots would provide the means for the distribution and amplification of these messages. Russian actors used networks of tens of thousands of automated accounts to artificially increase the visibility of their messages. When a piece of disinformation was posted, the botnet would instantly "retweet" and "share" the disinformation thousands of times within a matter of minutes (Dawson & Innes, 2019). The performance of these acts exploited the purpose of "Trending" algorithms on platforms like Twitter, which interpret rapid engagement as a sign of breaking news. By forcing a specific hashtag or narrative to trend, the botnet can bypass the "filter bubble" of individual users and push false narratives into the mainstream of the media cycle (National Institute of Justice, 2021). Journalists and politicians, who now, seeing a topic trending, would often comment on the topic which unknowingly lent credibility to a narrative that was created by a foreign intelligence service. This technique of "Artificial Amplification" through the use of bots, influences the heuristic processes of the human mind. Humans are evolutionarily programmed to pay attention to what the crowd is discussing. Bot networks emulate crowd sourced ideas which influence the thoughts of individuals.

4.4 Microtargeting and Ads

The final tactical layer of tactics employed by the IRA and GRU was the use of commercial advertising as tools for information warfare. Russian actors purchased over 3,000 political advertisements on Facebook and Instagram, spending roughly $100,000, which is a trivial sum to a state actor, but made highly effective due to microtargeting of individual audiences (Young & Kim, 2018). Unlike television ads, social media ads are not broadcast to a general audience but can be "microtargeted" to specific subsets of a population. Russian operatives used Facebook's ad tools to target users based on "Likes," location, and demographic (Zuboff, 2019). They crafted specific messages for African American voters intended to discourage turnout, while

simultaneously targeting conservatives with messages designed to incite fear regarding subjects such as immigration. The use of microtargeting as an application of propaganda exploits the “surveillance capitalism” model. The platforms provided the weapon (targeting tools) and the ammunition (user data) to Russian actors, who then pulled the trigger. This is also a failure of “Know Your Customer” (KYC) protocols in the digital ad market. In this case foreign actors can anonymously purchase influence in a domestic election using foreign currency and accomplish their goals with zero scrutiny (Quigley, 2017).

4.5 Russia Ukraine War

The Russian invasion of Ukraine showed a clear change in information warfare. Operations went from a secret, denial-based method used in 2016 into an open and complete control of information. The Internet Research Agency’s (IRA) early tactics slowly infiltrated natural political movements. The 2022-2025 strategy used a high-volume, high-pressure flow of falsehoods. Artificial intelligence and encrypted platforms boosted this strategy to support military attacks. A key tactical step up involved using synthetic media directly within a combat zone. In March 2022, Russian hackers took over the website of the Ukrainian news broadcaster Ukraine 24. They streamed a deepfake video of President Volodymyr Zelenskyy. In the video a fake Zelenskyy told Ukrainian troops to drop their weapons and surrender to the Russian armed forces. People quickly exposed the video as fake because of visual problems, such as a still body and a voice that did not match the lips. This video acted as a "Proof of Concept" for combining online influence operations with real military aims. It pushed social media weaponization past simply "sowing discord." It aimed to reach actual military goals, specifically getting enemy soldiers to surrender, using digital tricks.

5. Anticipatory Ethics

The failure to prevent the weaponization of social media was not only an unavoidable consequence of technology, it was a failure of foresight about social media vulnerability. Anticipatory Ethics is a framework which focuses on the development of potential ethical issues relating to new and emerging technologies before the technologies become fully entrenched (Brey, 2012). Applying the "Moral Responsibility for Computing Artifacts" rules by Miller et al. reveals the extent of the failures of social media platforms to foresee ethical issues due to the lack of content moderation.

5.1 Rule 4: The Sociotechnical Imperative

The Sociotechnical Imperative states that “People who design, develop, or deploy a computing artifact can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artifact is embedded” (Miller et al., 2011). Social media developers designed their tools for a sterile, idealized world where users are all believed to be authentic and who wish to make connections with others in an inherently positive way. The developers have entirely failed to take into account the sociotechnical reality of geopolitics, where bad actors exist and where information can be weaponized by these bad actors. They ignored the “Dark Scenarios” described by Philip Brey, failing to model how a state adversary can and will abuse their systems (Brey, 2012). By developing their algorithms solely for optimization of engagement, they basically built a machine that amplifies disinformation, hateful speech, and bad conduct. A responsible design process would have anticipated the vulnerabilities that could be exploited by bad state actors and they could have built “friction” into the system to slow the spread of unverified information. The failure to uphold the Sociotechnical Imperative has resulted in platform architecture that is vulnerable to PSYOPS employed in cognitively oriented cyberwarfare.

5.2 Rule 1: The Foreseeability of Effect

Rule 1 for computing artifacts asserts that “The people who design, develop, or deploy a computing artifact are morally responsible for that artifact, and for the foreseeable effects of that artifact” (Miller et al., 2011). The interference in the 2016 election was not an unforeseeable effect of the platforms design. It was known that open ad platforms allowed for anonymous purchasing. It was also known that algorithms prioritizing outrage can lead to radicalization. It was known that bots can mimic humans. The platforms cannot claim ignorance, since the vulnerabilities were features of how the social media platforms work, not anomalies. The adherence to a “move fast and break things” ethos of these social media companies prioritized speed and growth of the platforms over safety. This negligence of safety concerns renders the architects of the social platforms morally responsible for the consequences of the cognitively oriented warfare waged on their networks. They built the battlefield and sold the weapons, no different than the “Merchants of Death” of the late 19th and early 20th centuries, firing no shots themselves but causing irreversible social and political damage.

5.3 Rule 5: Honesty in Promotion

Rule 5 declares that “People who design, develop, or deploy a computing artifact must not deceive the users about the artifact” (Miller et al., 2011). Social media platforms market themselves as tools for “connection,” “community” and “giving people a voice.” This framing is incredibly deceptive. The reality is that the platforms functioned as surveillance engines that sold user attention to the highest bidder. By concealing the mechanics of the algorithms and the extent of data mining, the platforms deprived users of informed consent. Users believed they were engaging in a public square, not knowing that they were arguing against Russian bots and the “news” they were reading was targeted propaganda (National Institute of Justice, 2021). This deception prevented users from properly evaluating the information they consumed, making them vulnerable to psychological manipulation. Honesty in the promotion of their products would require social media platforms to clearly label automated accounts, disclose the source of advertisements, and explain why content is being recommended.

6. Recommendations

To counter the “Velocity,” “Virality,” and “Microtargeting” of modern cognitive cyberwarfare, by relying on human moderation is insufficient. The “OODA Loop” of a disinformation campaigns operates at machine speed, while democratic deliberation runs at the speed of human thought. Therefore, we propose a shift to Cognitive Cybersecurity, a defense architecture designed to secure the information environment through transparency and technical provenance.

The first pillar of this architecture is Cryptographic Content Provenance. Russian disinformation campaigns rely on the ability to pass off fabricated content from fake news sites or doctored images as authentic. To address this, we recommend the universal implementation of the C2PA (Coalition for Content Provenance and Authenticity) standard. This utilizes public key cryptography to digitally sign media at the source of its creation. When a journalist or government entity publishes a story, it is “stamped” with a cryptographic hash. Social media platforms could then display a “Verified Source” badge based off this data. If a Russian operative copies the image and alters anything, the hash breaks, and the platform automatically flags it as “Manipulated Media” (Rosenthal, 2021). This moves the burden of trust from the user’s judgement to a cryptographic infrastructure.

The second pillar involves Algorithmic “Circuit Breakers.” Disinformation exploits the speed of social media algorithms, where outrage spreads faster than fact checks can be produced. To counteract this, platforms must implement viral circuit breakers, a concept borrowed from financial markets where trading is halted during bouts of panic selling. This occurs if a piece of content exhibits “Anomalous Viral Velocity,” such as a sharing speed more than three standard deviation points above the normal, combined with low credibility flags (e.g., a newly created account with zero outbound links). This introduces “Friction” into the system, buying time for an automated fact checking system or human reviewer to assess the content before it reaches critical mass.

The third pillar addresses the issue of automated botnets through Zero-Knowledge Proof of Personhood. The IRA used botnets to create the illusion of a consensus in audiences, called Astroturfing. To neutralize this, we propose Proof of Personhood protocols to verify if a user is human without revealing their identity. Using Zero-Knowledge Proofs (ZKPs), a user can prove they are a unique human via a decentralized identity token without sharing their name or government ID with the platform. Accounts that lack this “Humanity Token” would be rate-limited or frozen, preventing them from mass-retweeting or mass-commenting. This effectively decreases the influence of botnets while preserving the user’s right to privacy (Tavani, 2009). Finally, we recommend an Algorithmic Transparency API to allow vetted researchers to audit recommendation engines, ensuring integrity against foreign agents.

7. Future Work

While this analysis has focused on the tactics of the 2016-2022 era, which were characterized by human controlled troll farms and manual coordination, the future of cognitive cyberwarfare lies in the automation of these operations using Generative Artificial Intelligence (GenAI).

Current Russian doctrine relies on the Internet Research Agency (IRA) employing humans to write the divisive comments. Future research must investigate the rise of "Agentic Disinformation Systems." As noted by Pauwels (2024), we are moving from human-driven troll farms to LLM-driven autonomous agents capable of engaging in sustained, context-aware debates with thousands of users simultaneously.

This shifts the bottleneck of propaganda from "human labor" to "compute power," allowing for a scale of manipulation previously impossible. Human trolls are limited by fatigue and language barriers, LLMs can generate infinite, linguistically sound, and contextually relevant propaganda at zero marginal cost. A single "Agentic AI" could replace an entire floor of human operators, constructing thousands of arguments at a time in the comment sections of their target audiences (Pauwels, 2024).

The next frontier of disinformation is Synthetic Media. As Generative Adversarial Networks (GANs) become accessible, state actors will move beyond mere text to audio and video fabrication. Future studies should analyze the impact of "deepfakes", synthetic audio and video clips released on private messaging apps (such as WhatsApp or Telegram) where attribution is impossible and viral spread is end-to-end encrypted. We anticipate a shift from "Broadcasting" fake news to "Narrowcasting" fake reality, where specific individuals will be targeted with deepfake voicemails and videos from trusted contacts (e.g., boss or family member) designed to elicit immediate behavioral compliance (Helmus, 2022).

Finally, future research must address the threat of Data Poisoning. As western nations grow more reliant on AI analysis for decision making, Russian actors are likely to switch from influencing people to influencing the Large Language Models themselves. By flooding the open web with subtle disinformation, they aim to "poison the well" by poisoning the data on which AI models are trained, in the attempt to ensure Western AI systems adopt a pro-Russian bias, or to spew historical revisionism (Schneier, 2023).

8. Conclusion

The Russian weaponization of social media shows that cognitive cyberwarfare is a present and escalating reality. The tactics of the IRA which employs disinformation, troll farms, microtargeting, and bot networks, are not trivial nuisances. They are state-sponsored psychological operations designed to degrade the sovereignty and autonomy of nation states. This form of warfare exploits the vulnerabilities of the human mind and architecture of the internet to achieve strategic objectives without the need for kinetic attacks

By applying the frameworks of standard ethical principles, the ACM code of ethics and Anticipatory Ethics, we conclude that the primary vulnerability lies not with the people, but with the platforms on which cognitive cyberwarfare takes place. The prioritization of engagement over truth, commodification of user data, and the lack of verification have created a "security hole" in the democratic process that the Russian operatives exploited. The failure of platforms to uphold the Sociotechnical Imperative and the principle of Foreseeability of Effect represent a profound ethical lapse in judgement.

To defend against these threats, we need to move beyond "content moderation" which is a reactive process. We must implement a posture of Cognitive Cybersecurity. This requires a redesign of social media architectures to prioritize authenticity over anonymity, friction over virality, and transparency over opacity. It requires the implementation of defensive AI to detect a coordinated threat by its behavior and the establishment of ethical standards that treat the attention of a user as a protected asset. The defense of the "Human Element" in the digital age is no longer a matter of internet literacy; it is a matter of engineering. It is the moral obligation of the computing profession to build digital infrastructures that are protected from manipulation, which will then protect rather than exploit the cognitive literacy of the citizens that inhabit the cyber sphere.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.0 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing. Gemini 3.0 pro used for aiding in sourcing.

References

Almond, D., Du, X., & Tang, A. (2022). Reduced trolling on Russian holidays and daily U.S. presidential election odds. *PLOS ONE*. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0264507

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. https://www.acm.org/code-of-ethics

Atlantic Council. (2022). *Deepfakes in the War in Ukraine: A Threat Assessment*. Digital Forensic Research Lab (DFRLab).

Bipartisan Policy Center. (2023). *The Pros and Cons of Social Media Algorithms*. https://bipartisanpolicy.org/wp-content/uploads/2023/10/BPC_Tech-Algorithm-Tradeoffs_R01.pdf

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Claverie, B., & du Cluzel, F. (2022). Cognitive Warfare: The New Battlefield Exploiting Our Brains. *Polytechnique Insights*. https://www.polytechnique-insights.com/en/columns/geopolitics/cognitive-warfare-the-new-battlefield-exploiting-our-brains/

Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press.

Dawson, A., & Innes, M. (2019). How Russia's Internet Research Agency Built its Disinformation Campaign. *The Political Quarterly*, 90(2), 245–256.

Digital Forensic Research Lab (DFRLab). (2024). *Doppelganger: How Russia mimicked real news sites and created fake ones to target US audiences*. Atlantic Council. https://dfrlab.org/2024/09/18/doppelganger-us-election/

Google Threat Intelligence Group. (2025). *Pro-Russia Information Operations Leverage Russian Drone Incursions into Polish Airspace*. Google Cloud Blog. https://cloud.google.com/blog/topics/threat-intelligence

Helmus, T. C. (2022). *Artificial Intelligence, Deepfakes, and Disinformation: A Primer*. RAND Corporation.

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

NATO Allied Command Transformation. (2021). *Cognitive Warfare*. https://www.act.nato.int/activities/cognitive-warfare/

NATO Allied Command Transformation. (2024). *Cognitive Warfare: A Strategic Concept*. NATO ACT. https://www.act.nato.int/activities/cognitive-warfare/

National Institute of Justice. (2021). *The Role of the Internet and Social Media on Radicalization*. https://www.ojp.gov/pdffiles1/nij/305797.pdf

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

Pauwels, E. (2024). *The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI*. United Nations University Centre for Policy Research.

Quigley, M. (2017). *Vanity Fair: Did Jared Kushner's Data Operation Help Select Facebook Targets For The Russians?* https://quigley.house.gov/media-center/news-article/vanity-fair-did-jared-kushner-s-data-operation-help-select-facebook

Royal United Services Institute (RUSI). (2025). *Russia, AI, and the Future of Disinformation Warfare*. RUSI. https://static.rusi.org/russia-ai-and-the-future-of-disinformation-warfare.pdf

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The Weaponization of Social Media*. Eamon Dolan/Houghton Mifflin Harcourt.

Tavani, H.T. (2009). *Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology*. Wiley.

Tech Policy Press. (2025). *Ukraine's Hard-Won Approach to Strategic Communications and Counter-Disinformation*. Tech Policy Press. https://www.techpolicy.press/ukraines-hardwon-approach

Thomas, E. (2024). *Russian influence operation Doppelganger linked to fringe advertising company*. Institute for Strategic Dialogue (ISD). https://www.isdglobal.org/digital_dispatches/russian-influence-operation-doppelganger

United States Department of Justice. (2024). *Justice Department Seizes 32 Domains Used in Russian Government-Directed Foreign Malign Influence Campaigns*. Office of Public Affairs. https://www.justice.gov/opa/pr/justice-department-seizes-32-domains-russian-influence

United States Department of State. (2024). *Disarming Disinformation: Russia's Use of Chemical Weapons Narratives in Ukraine*. Global Engagement Center. https://2021-2025.state.gov/disarming-disinformation/

Young, M., & Kim, Y. (2018). *Uncover: Strategies and Tactics of Russian Interference in US Elections*. University of Wisconsin–Madison.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Echo Chambers, Filter Bubbles, and Cyber Warfare: An Ethical and Anticipatory Ethical Analysis

Richard Wilson[1,2] and Noah Donnelly[2]
[1]Department of Philosophy, Towson University, Baltimore, Maryland, USA
[2]Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA

wilson@towson.edu
ndonnel1@students.towson.edu

Abstract: With the development of the internet in the information era and the wide access to information the internet makes available, Echo Chambers and Filter Bubbles have developed. Echo Chambers and Filter Bubbles are a consequence of Reinforcement Learning Algorithms. An Echo Chamber is an environment where people only encounter beliefs or opinions that reinforce the beliefs and opinions to which they are already committed. (Sunstein, 2017). This serves the purpose of creating a constant positive feedback loop, which continually reinforces one idea or set of ideas (Murphy, 2022). A Filter Bubble develops when recommendation algorithms feed users content based on what narratives the recommendation algorithm determines an audience wants to hear to maximize user engagement (Pariser, 2011). Echo Chambers and Filter Bubbles are relevant to Cyber and Cognitive Warfare because state actors can take advantage of the existence of these environments and use them to influence discourse and public opinion. In the context of cyber warfare attackers can use bots and fake accounts where they pretend to be citizens of the target nation to flood these Echo Chambers with narratives that align with the audience's current belief system and while also benefiting the attackers (Singer & Brooking, 2018). In the context of cyber warfare attackers can abuse Filter Bubbles by using data breaches and advertising data to infiltrate the Filter Bubbles and direct them towards the attackers desired narrative (Matz et al., 2017). These abuses rise above the level of mere internet trolling, they are intentional and targeted acts of Cyber Warfare aimed at influencing the cognitive space of a nation's population (Claverie & du Cluzel, 2022). Echo Chambers and Filter Bubbles are constructed to accomplish a strategic objective, in this case they are aimed at influencing the decision making and opinions of an audience to influence elections, policy, protests, or overall public sentiment. These strategic objectives are accomplished by using non-kinetic weapons to destabilize society, diminish decision making capabilities, and erode trust in democratic institutions. This type of cyber-attack was made apparent in the COVID-19 disinformation schemes when wellness communities on social media platforms were flooded with anti-vaccine and medicine narratives leading to distrust in the medical system, and the politicians who promoted them (Dawson & Innes, 2019). This paper identifies the technical, ethical, and anticipated ethical issues of Filter Bubbles and Echo Chambers and proposes a technical and policy framework to classify and prevent these acts of Cyber Warfare.

Keywords: Echo chambers, Filter bubbles, Cognitive warfare, Information operations, Algorithms, Micro-targeting, Anticipatory ethics, Epistemic paternalism, OCEAN model

1. Introduction

In the domain of modern conflict, perhaps the most potent form of warfare is the cognitive domain. Social media platforms and content recommendation algorithms have led to the development of two distinct yet related phenomena: Echo Chambers and Filter Bubbles. Echo Chambers are online areas where a person only encounters a point of view or opinion that mirrors their own opinions. Filter Bubbles endorse single track mindsets created by search algorithms calculating what the user wants to hear or see based on information collected about them such as: location, past-click behavior, and search history (Pariser, 2011). Initially echo chambers and filter bubbles were viewed as sociological side effects of the information era, but in truth they have now been weaponized by malicious state actors as a means of cyber warfare.

The combination of echo chambers, and filter bubbles with cyber warfare attacks represents an evolution in Information Warfare (IW). Traditional cyber warfare focused on stealing data and destroying or manipulating critical infrastructure. Echo chambers now represent a new paradigm, they focus on destroying social cohesion, and the manipulation of truth making a popular consensus on social topics difficult to attain. Attackers can insert disinformation into targeted echo chambers which can cause the public sentiment about an adversary nation to change without the use of any kinetic force. This can be called Non-Kinetic Warfare which targets the ability of a nation's people to make informed and educated decisions. These tactics can destabilize societies, influence elections, and erode trust in their democratic institutions, which align perfectly with the goals of authoritarian and fascistic styles regimes (Mercy Corps, 2020).

This paper argues that echo chambers and filter bubbles are being manipulated to obscure facts and truth and the diversity of opinion and they should be recognized and classified as a form of cyber warfare, which requires a new ethical framework for assessment. On the technical side we will analyze the mechanisms which are used

to manipulate algorithms and micro-target populations. On the ethical side we will investigate the effects and implications of cognitive autonomy being eroded and apply the principles of Anticipatory Ethics to identify the moral responsibilities of platform owners. By analyzing case studies which include, Cambridge Analytica, the Myanmar genocide, and the Russian use of botnets, we show that the mode of attack in the 21st century is the human mind and not just a kinetic military base.

2. Technical Issues

Recognizing the weaponization of echo chambers and filter bubbles requires the examination of the technology that allows and facilitates their existence. The technical analysis provides the foundation for ethical and anticipatory ethical analysis of echo chambers and filter bubbles. Influencing vulnerable target audiences is an intended function of the recommendation algorithms and the data structures they rely on.

2.1 The Mathematics of Radicalization

Social media platforms use complicated Machine Learning (ML) algorithms, mainly Reinforcement Learning (RL) which is intended to maximize user engagement, to target users. The algorithms that are employed use a tactic called Reward Function, where success is measured by time-on-site or interaction (clicks, likes, shares). The algorithm observes the user's state (current mood, history), performs the action of recommending content, and receives a reward in the form of user engagement. Algorithms have learned that high-arousal emotions such as fear, outrage and validation generate the highest engagement rewards (Vosoughi et al., 2018). This creates a Positive Feedback Loop for target audiences. The algorithm feeds users increasingly extreme content to achieve higher engagement rewards, which in turn pushes the user deeper into the filter bubble, which can lead to radicalization. Radicalization can be exploited through injecting content that is mathematically optimized to trigger the highest engagement rewards, which can at the same time be radical disinformation. Botnets can artificially inflate engagement metrics of extremely divisive content, effectively gamifying the Reinforcement Learning (RL) model, which tricks the RL model into promoting disinformation to millions of real users. The algorithm is unable to recognize truth, so it becomes an unwitting technological accomplice in the radicalization of target audiences

2.2 Psychometric Micro-Targeting and the OCEAN Model

Algorithmic attacks using echo chambers, and filter bubbles are extremely precise due to the large amount of personal data on social media platforms, typically gained through data brokers or data breaches. The large amounts of data gathered about individuals allow for Psychometric Micro-targeting. This technique transcends demographics (age, location) and instead targets personality traits. This is where the OCEAN Model (McCrae & John, 1992) can be applied, and the so called Big Five personality traits can be targeted: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Users of social media have a "digital exhaust" which are recorded as likes, shares and browsing history. Algorithms can predict the traits associated with the digital exhaust with a high degree of accuracy. If an attacker identifies a user who scores high on the Neuroticism scale which would make them prone to anxiety and fear, they can be targeted by disinformation employing narratives that misinform them about looming threats. An important category of psychometric micro-targeting is users with low Openness, who are targeted with more traditional propaganda. This transforms propaganda from a blunt, generalized tool into a more precise, sharp tool tailored to specific neurological vulnerabilities of a micro-targeted audience (Matz et al., 2017).

2.3 Bot-Human Hybrid Network

"Sybil Attacks" are disinformation campaigns wherein a single adversary controls multiple fake identities to manipulate the trust network metrics. (See: Gunturu, Rupesh) Modern operations use scripts from fake accounts and AI-driven personas, so when someone engages in an argument online, it may not even be with another human. Russian botnets have begun to integrate Large Language Models (LLMs) into botnets, generating contextually relevant comments that are unique, but which are created by bots. These bots can argue for prolonged periods of time within online comment sections and online forum threads which gives the appearance of a public consensus. This technique is referred to as AstroTurfing. Attackers can map the Social Graph of an echo chamber to identify Bridge Nodes which are users who have influence and who span multiple communities. By compromising or controlling Bridge Nodes attackers can infect multiple echo chambers across multiple platforms causing the misinformation presented in the echo chamber to cascade (Ferrara et al., 2016; Tomassi et al., 2024).

3. Ethical Issues

The weaponization of echo chambers and filter bubbles raises important ethical questions regarding individual autonomy and what responsibilities states have for protecting individual autonomy.

3.1 Cognitive Sovereignty and Consent

The central ethical violation in this kind of warfare is related to what is known as Cognitive Sovereignty, (Claverie & du Cluzel, 2022). Cognitive sovereignty is the right of an individual to control their mental processes and for that individual to be free opinions influenced by covert operations. When users sign up for social media, they consent to receiving targeted advertisements, but there is no clause stating that they consent to being the victim of state-sponsored psychological operations (PSYOPS). Hidden algorithms being used to curate people's perceptions of reality deprives them of the agency to make informed decisions. This violates Kant's 2nd Categorical Imperative which states that one must treat humanity as an end-in-itself, as opposed to a political end. Users become potential puppets for foreign governments in a conflict that they are unaware of, where their perception of reality has been altered by foreign adversaries (Claverie & du Cluzel, 2022). The question is to what degree states and social media platforms have to protect the cognitive sovereignty of individuals.

3.2 Epistemic Paternalism and the "Truth" Dilemma

Countering these attacks raises the ethical issue of Epistemic Paternalism. Platforms are effectively in a "Catch 22" dilemma where if they break the filter bubbles and downrank disinformation, they are now the arbiters of what is "true" or 'false" or "good" or "bad" in the world and of what the users should see. Corporate algorithms become the chief judge of what is true and what is good information. Contrarily if the platforms do not intervene, cyber warfare is now capable of destabilizing democracy. If a social media platform does intervene, it is now acting as a paternal figure to the user, determining the information audiences are and are not allowed to see. There is not a neutral setting for these platforms, so what must be determined is what causes the least harm, and who has the moral authority to regulate these values: states, corporations, or users? (Aird, 2022).

3.3 The Erosion of Trust and Democratic Norms

Echo chambers and filter bubbles channel citizens towards different and often conflicting realities, making consensus between opposing groups nearly impossible. From a Utilitarian outcomes-based standpoint, this causes an incredible amount of harm to the stability of democratic societies, allowing democracy to erode. This channeling of citizens towards different and often conflicting realities has contributed to the erosion of trust in democratic institutions. When a population cannot agree on the most basic of facts like the effectiveness of vaccines or the legitimacy of an election, the social contract between citizens and fellow citizens and between citizens and government has broken down. Cyber warfare conducted through echo chambers and filter bubbles targets the trust of the people in the institutions of a nation, which is just as critical for a nation as its electrical grid. Unlike physical infrastructure, once this type of trust is broken, it is difficult to rebuild.

4. Case Studies

In the early stages of the development of the internet the dangers of echo chambers were previously largely theoretical, their exploitation was not yet well known, however, they recently have materialized and have shown themselves to be to date one of the most dangerous non- kinetic weapons. The following case studies will demonstrate how state and non-state actors have used the technical architecture of social media platforms to their advantage when undergoing strategic military and political operations employing echo chambers and filter bubbles.

4.1 Cambridge Analytica (2016)

Cambridge Analytica is essentially the founding father of psychometric profiling and of enabling the ability for bad actors to use psychometric profiling as a means for carrying out cognitive warfare. Cambridge Analytica collected the personal data of roughly 87 million Facebook members without the users consent through a third-party quiz app named "This Is Your Digital Life." This was not simple demographic targeting, but a much more personalized means of targeting audiences. The engineers then fed data into the OCEAN Model (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) to psychologically profile every voter in the United States. Through an analysis of what was dubbed a user's "digital exhaust" that included likes, shares and comments, Cambridge Analytica's algorithms were able to predict personality traits with such accuracy that they overrode the acceptance of the opinions of users own family members. The firm was able to identify "Persuadables" in swing states, who were users who scored high in Neuroticism and low in Openness. The

identified users were then fed "dark posts" which were nonpublic advertisements only visible to the target audiences. These dark posts were mathematically engineered to trigger specific anxieties and fears in audiences related to immigration or government overreach. This was an exploitation of filter bubble architecture, Cambridge Analytica helped make fictive personalized realities for millions of voters based off their deepest vulnerabilities. This was the great hack of the democratic decision-making process, voting machines were no longer the target, it was the voters mind (Cadwalladr, 2018).

4.2 Myanmar Genocide (2017)

In Myanmar**, Facebook's engagement-based recommendation** algorithms were abused by military operatives to spark a genocide against the Rohingya Muslim minority. This is the direct, kinetic consequence of algorithmic amplification. The military of Myanmar (Tatmadaw) launched a secret operation using fake accounts and "sock puppet" pages that were disguised as entertainment or celebrity news to enter into, infiltrate and influence the digital lives of the people they were supposed to protect. Once these accounts were integrated into the user's daily lives, they started to spread anti Rohingya Muslim propaganda labeling the Rohingya population as an existential threat to Buddhism. The technological failure lay in the hands of Facebook's engagement ranking algorithms, which had the consequence of prioritizing hate speech since it generated intense engagement, while outrage and fear scored the highest in "time-on-site" metrics. The platform then became an enormous, radicalized echo chamber, and calls to violence were subsequently boosted by the algorithm and spread to millions of users who had no alternate news sources. This digital incitement led to physical violence, leading to the murder of at least 10,000 Rohingya people and the displacement of over 700,000 of them. This is proof of flooding filter bubbles, filled with state-actuated hate speech, which resulted in more than polarization. In the Myanmar Genocide case, the platforms code aided in a mass atrocity (Mozur, 2018).

4.3 Russian Doppelganger Operation (2022-Present): The Clone War

During February of 2022 Russia began their ongoing invasion and war against Ukraine. Since that time they had also begun a sophisticated campaign known as "Doppelganger" to destabilize the European support for Kyiv. This operation uses a "Supply Chain Attack" on truth by creating clones of legitimate Western media. Russian operatives set up domains that would try to prey on typos. They set up e. g. bild.ltd instead of bild.de, and theguardian.co.com instead of theguardian.com, and flooded these sites with fake articles (with disinformation) that mimicked the visual style including fonts and bylines of the original publications and legitimate sources. These fake articles advocated narratives that argued sanctions on Russia were destroying the European economy and that Ukrainian refugees were committing heinous crimes. If they had just deployed these websites they would gain very little traction, so in addition they had to use their botnets to flood social media sites with comments and posts that are linked to these fake sites. The sophistication of this attack lay in the ability to bypass an individual user's skepticism filter, which creates a situation where a user who trusts The Guardian, is now likely to trust these fake sources since the disinformation presented to the audience looks like it comes from a legitimate real source. By targeting very specific filter bubbles, such as French farmers who are worried about fuel prices, or people working in the German manufacturing sector that feared layoffs, the Doppelganger operation feeds false enemy narratives straight into domestic political conversations of NATO nations, which exploits the trust of a targeted population in the free press (US Cyber Command, 2024).

4.4 COVID-19 Disinformation

During COVID-19, there was a concerted effort to radicalize new audiences by state-actors. These bad actors started targeting the "Wellness" communities filter bubbles to radicalize a new audience though the "Pastel QAnon" phenomenon. Disinformation campaigns began to target communities that were focused on yoga, organic foods, and alternative medicines, groups that already harbored a latent distrust of the pharmaceutical companies and government mandates coming from government officials. State-actors and domestic extremists used the algorithm to connect users interested in "natural immunity" with groups promoting "medical freedom," which served in turn as the gateway to the darker conspiracy theories that talked about "The Great Reset" and QAnon messaging about globalist cabals. Russian and Chinese state actors amplified these messages to undermine Western public health responses from government officials. The "Data Void" technical exploit was where attackers generated content for unique search terms like "Plandemic" to make sure their conspiracy theories dominated search results well before they could be fact checked and before authoritative sources could respond. The wellness communities, which were previously untouched by far-right political extremism were now effectively pipelined to a point where reality became secondary to political disinformation narratives. This materialized in real world harm that included refusing vaccines, harassing medical professionals, and destabilizing public health infrastructure related to trusting healthcare information coming from government

officials. This case shows how any filter bubble can be abused, even ones focused on health and positive information, for the purposes of political extremism and destabilizing a political enemy (Claverie & du Cluzel, 2022).

5. Anticipatory Ethics

By applying Anticipatory Ethics to these case studies, what gets discovered are the moral obligations of developers of social media platforms and lawmakers related to echo chambers and filter bubbles before this next generation of warfare is fully deployed. Anticipatory ethics is a recent development in ethics concerned with examining ethical issues with technologies and technological artifacts from the research and development stage, through the introduction stage, to the stage of marketplace permeation and saturation. (Brey, 2012) Anticipatory engineering ethics and anticipatory ethical analysis attempts to identify ethical issues with Engineering and ICT technologies before and as technology develops and as they are introduced. This type of ethical analysis can also be extended to technologies related to conducting cyber warfare. (Wilson, 2021). This section will analyze these conflicts through the lens of "Moral Responsibility for Computing Artifacts" rules developed by Miller et al. (2011) and the Association for Computing Machinery (ACM) Code of Ethics.

5.1 Rule 1: The Foreseeability of Effect and ACM Principle 1.2

Rule 1 of Anticipatory Ethics employing Moral Responsibility for Computing Artifacts states that "The people who design... are morally responsible for that artifact, and for the foreseeable effects of that artifact" (Miller et al., 2011). When applied to echo chambers and filter bubbles, it is undeniably foreseeable that algorithms that are rewarded for engagement no matter the type of message, could be used by state-actors to radicalize users. The typical argument made by the companies that own the platforms who try to spin themselves as neutral town squares, is ethically invalid when the town square is designed to reward the most toxic and divisive content that they host, all in the name of “time-on-site” metrics. This design choice violates ACM Principle 1.2 “Avoid Harm” (Association for Computing Machinery, 2018). Harm is not just a metric of physical injury; it is also psychological violence. Psychological violence and the destruction of social cohesion, the incitement of violence against a specific ethnic group, and the degrading of democratic voter’s abilities to make informed decisions are negative consequences. Anticipatory Ethics requires developers to move beyond reactive content moderation, which allows harm to spread, but now mandates them to implement “Red Teaming” for cognitive risks during the design phase of their technologies. This would involve simulating how bad state actors, such as the Russian Internet Research Agency, think and operate. Through this type of analysis, they could secure a new feature they are thinking about rolling out, before it is released to the public. Failing to anticipate these potential exploitations, given the amount of undeniable evidence about previous election interferences, represents a failure of their professional duty and a breach of the social contract that exists between consumer and producer as well as between producer and government.

5.2 Rule 4: The Sociotechnical Imperative and ACM Principle 3.7

Rule 4 tells us that "People who design... can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artifact is embedded" (Miller et al., 2011). The sociotechnical systems were designed for developing friendships, connections, and sharing. These platforms are currently operating in systems defined by “information warfare,” “geopolitical conflict” and “hybrid threats.” By not adapting their architecture to this new reality of hostile and volatile internet occurrences, there is a failure to uphold the Sociotechnical Imperative. This corresponds with ACM Principle 3.7 where professionals are required to "Recognize and take care of systems that become integrated into the infrastructure of society" (Association for Computing Machinery, 2018). Since social media has become the new public sphere where information is spread, including information about health policy, elections, and debates, developers should now carry a burden related to the duty of care. A responsible developer would implement algorithms that by design expose users to a diverse array of viewpoints, as opposed to the current algorithms that promote tunnel vision through echo chambers and that reinforce filter bubbles. When Social media companies allow their deployed algorithms to remain this way, the developers are essentially splitting the population into feudal warring tribes on a digital stage instead of a traditional battlefield. Their algorithms are inherently counterintuitive to a functioning democratic system.

5.3 Rule 2: The Post Deployment Mandate and the Problem of Epistemic Paternalism

Rule 2 tells developers that "Responsibility includes being answerable for the behaviors of the artifact... after deployment" (Miller et al., 2011). Platforms duties do not end at deployment; they have a continuing duty to

monitor their interpersonal ecosystems and moderate them. They cannot abdicate responsibility at the time of deployment. However, by strictly enforcing this mandate they introduce the issue of Epistemic Paternalism, which is the idea that platforms must intervene in the user's acquisition of knowledge "for their own good" (Aird, 2022). A counterargument posed to this could be if a platform breaks someone's filter bubble or scrubs disinformation, they act as an ultimate authority who decides what is true and what is not. This limiting of the autonomy of the user can be argued to be like how we treat children and animals, not other human beings. However, while limiting autonomy is considered unethical, an Anticipatory Ethical analysis could argue that if the user's autonomy has already been compromised by foreign state actors, engaged in attacking the platform user's psyche, the platform is morally obliged to intervene. In this context, any intervention is not to control what a user thinks, but to restore their ability to think autonomously, more specifically, to restore the user's ability to be able to think freely without interference from covert manipulation. This would morally justify aggressive actions like banning botnets, labelling state-controlled media, and friction induced sharing prompts not as acts of censorship but as a necessitated "Cognitive Defense."

6. Recommendations

To eliminate the threat of cyber warfare that utilizes echo chambers and filter bubbles, the over-reliance on passive content moderation must be abandoned, and in its place, a multi-layered defensive strategy that involves technical reforms, re-engineering of algorithms, and cognitive security protocols needs to be developed.

6.1 Diversity Aware Recommendation Systems (DARS)

The technical aspect of the target of recommendation systems used for radicalization was the "Similarity Bias" that is at the center of collaborative filtering environments. To defend against this, platforms need to implement Diversity-Aware Recommender Systems (DARS) that use Determinantal Point Processes (DPP). Basic matric factorization models optimize only for relevance (predicting what a user is going to click). In contrast to this, a DPP-based layer would model a diverse set of items, that would ensure that the user is fed a broader semantic space as opposed to being trapped in their filter bubble. We propose that the objective function of the recommendation engine should include a "Serendipity Hyperparameter" (λ). This parameter is meant to punish semantic redundancy. If the vector embedding of the top five recommended articles are the same, it indicates identical political narratives, and the DPP layer would then introduce a repulsive force, switching redundant articles for content that is semantically distinct and highly rated by a diverse set of user clusters. This proposition would mathematically enforce the rule that people should be exposed to conflicting narratives, allowing them autonomy to choose which narratives they think are best, popping their filter bubble without requiring slow human moderation (Helberger et al., 2018).

6.2 Adversarial Training and Inverse Propensity Scoring

Adversaries attack Reinforcement Learning (RL) systems when they inject "poisoned" data into a system, which are fake likes, fake shares, and fake comments aimed at skewing the reward function. To eliminate this, platforms need to strengthen their Reinforcement Learning models with Adversarial Training. This would involve subjecting the RL model to training that would allow it to "Red Team" itself to eliminate recommendations that are likely to lead to radicalization. Platforms must also implement "Inverse Propensity Scoring" (IPS) which detects and discounts engagement metrics that are boosted by inauthentic behavior patterns, like those of bots. We recommend a "Trust Weighted Reward Function" where the platform multiplies the value of a signal such as a like or share by the users Reputational Scores which are determined by their network centrality and the age of their account. If a piece of content is being amplified by a cluster of low reputation accounts, the algorithm gives it a negative weight which suppresses the disinformation. This neutralizes Astroturfing campaigns by allowing only verified, and organic humans can be allowed to drive what is trending on a platform at any given time (Hussain et al., 2025; Platt et al., 2024).

6.3 Cryptographic Content Provenance (C2PA)

The "Doppelganger" operations and "Supply Chain Attacks" on truth must be identified and defeated. To overcome these attacks, the internet needs to adopt Cryptographic Content Provenance standards like C2PA (Coalition for Content Provenance and Authenticity). This standard would bind a cryptographically secure signature to all media files at the point of creation or publication. Respected news organizations and government agencies would be required to sign their legitimate content using their private keys. Social media platforms would then be required to implement trust indicators that act as digital watermarks and verify the chain of custody of that information. If the chain is broken, so is the digital signature, where it would lose its verified

status. So, if a Russian operative tried to clone The Guardian's website, it would be missing their signature, and the user would be alerted to the spoof. Since the user is already cognitively overloaded, the burden of verification is put upon the backs of the browsers SSL/TLS infrastructure which creates a chain of trust that state actors would be unable to create (Coalition for Content Provenance and Authenticity, 2023; Bushey, 2025).

6.4 Privacy-Preserving Algorithmic Auditing by Differential Privacy

Trust in the digital age requires independent verification, but the platforms that host their algorithms hide behind "trade secrets" and "user privacy" to prevent audits. We recommend the adoption of Differential Privacy interfaces for the purpose of auditing. This would allow independent researchers to query the social media platforms datasets to measure for the trajectory of radicalization of users without exposing any individual user's data. By adding controlled statistical noise to the query results, the platforms can prove to the regulators that their platforms are not pushing users towards extremism. This component solves the "Black Box" problem, it allows for auditing and enforcement of algorithmic liability laws without compromising their closed source software and user anonymity (O'Neil, 2016).

6.5 Cognitive Security Heuristics and Friction-by-Design

Cognitive Security (CogSec) is focused on hardening humans against emotional exploitation. Platforms must add "Friction-by-Design" heuristics. An example of this are Natural Language Processing (NLP) classifiers that are able to detect "High-Arousal/Low-Fact" linguistic patterns that are commonly found in disinformation publications. When a user attempts to share the newly flagged content without clicking on the source link, the system should throw them a "Read First" latency prompt. Platforms should also implement Inoculation Theory techniques where users are warned about disinformation tactics (e.g., "Warning: Users are currently targeting this topic with fake generated images") before they even engage with the feed. This forces users into an analytical thinking stage, breaking the outrage sharing loop (Maertens et al., 2024).

7. Future Work

Future research needs to analyze the escalation of cognitive weaponry that is driven by Generative AI. We predict the rise of "Hyper-Personalized Propaganda," where Large Language Models (LLMs) will be able to generate unique and persuasive messages for every single potential voter by scanning their real-time psychometric data, which would make "template matching" obsolete. Future work needs to explore Neuro-Symbolic AI defenses, which combine the pattern recognition abilities of neural networks and the logical reasoning of AI. Current models only understand word associations, Neuro-Symbolic AI could evaluate the factual consistency of a claim, allowing for automated fact checking. Additionally, research into Cognitive Immunology is required, specifically "AI Guardians" which are personal agents that run on a user's device to weed out manipulative content and highlight logical fallacies in real-time. This would act as a digital immune system and a potential defense for the mind of the user.

8. Conclusion

The manipulation of echo chambers and filter bubbles should be counted as a sophisticated form of cyber warfare, targeting not physical infrastructure, but now targeting the cognition of individuals instead. Through the analysis of technical mechanisms like Reinforcement Learning Loops, the OCEAN model, and botnet amplification, we demonstrate that these are not mere accidents, but engineered vulnerabilities that state-actors and radicals can use to cause harm while bypassing traditional military defenses. The case studies of Cambridge Analytica, the Myanmar genocide, and the Russian "Doppelganger" operations are proof that when content curation is left unguarded, the result is the incitement of violence with the additional result of degradation of democratic sovereignty.

This analysis also concludes that defense is in fact possible by developing strategies through the application of Anticipatory Ethics to the engineering process. Redesigning recommender systems to prioritize content diversity, enforcing Cryptographic Content Provenance, and requiring Trust-Aware Reward Functions, are all actions engineers can take to hedge these platforms against manipulation. The Sociotechnical Imperative refers to the pressing need to align technological development and social systems to achieve sustainability goals (See: , requires that developers must treat recommendation algorithms as critical infrastructure. We must now acknowledge that freedom of speech does not imply that an algorithm can recommend everything under the sun. To secure our democratic future, we need to defend the integrity of the human mind with the same rigor we apply to the defense of our military bases and outposts.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.1 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing.

References

Alabama Policy Institute. (2020, November). *Understanding the Difference Between Positive and Negative Rights*. Alabama Policy Institute.

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. https://www.acm.org/code-of-ethics

Aird, R. (2022). A puzzle of epistemic paternalism. *Philosophical Psychology*, *36*(4), 1011–1029. https://doi.org/10.1080/09515089.2022.2146490

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Bushey, J. (2025). Cryptographic provenance and AI-generated images. *AI Collaboratory*. https://ai-collaboratory.net/wp-content/uploads/2025/11/S13212_7356.pdf

Cadwalladr, C. (2018). *The Great Hack: The Scandal of Cambridge Analytica*. The Observer.

Claverie, B., & du Cluzel, F. (2022). Cognitive Warfare: The New Battlefield Exploiting Our Brains. *Polytechnique Insights*.

Coalition for Content Provenance and Authenticity. (2023). *C2PA technical specification*. https://c2pa.org/specifications/specifications/1.3/specs/C2PA_Specification.html

Collingridge, D. (1980). *The Social Control of Technology*. St. Martin's Press.

Dawson, A., & Innes, M. (2019). How Russia's Internet Research Agency Built its Disinformation Campaign. *The Political Quarterly*, 90(2), 245–256.

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM*, 59(7), 96-104.

Gunturu, Rupesh. Survey of Sybil Attacks in Social Networks. 1504.05522 (accessed 2/5/2026).

Hussain, M., Mehmood, A., Khan, M. A., Khan, R., & Lloret, J. (2025). Reputation-based leader selection consensus algorithm with rewards for blockchain technology. *Computers*, *14*(1), 20. https://doi.org/10.3390/computers14010020

Maertens, R., Spampatti, T., & van der Linden, S. (2024). Adding 'grit' to boost psychological inoculation against misinformation. *Journal of Experimental Psychology: Applied*. Advance online publication. https://osf.io/dzbm7/download/?format=pdf

Matz, S. C., Kosinski, M., Nave, G., & Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the National Academy of Sciences*, 114(48), 12714-12719.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, 60(2), 175-215.

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

Mercy Corps. (2020). *The weaponization of social media: Landscapes assessment and case studies*. https://www.mercycorps.org/sites/default/files/2020-01/Weaponization_Social_Media_FINAL_Nov2019.pdf

Mozur, P. (2018). *A Genocide Incited on Facebook, With Posts from Myanmar's Military*. The New York Times.

Murphy, T. F. (2022). *Feedback loops in psychology: Positive vs. negative loops for change*. Psychology Fanatic. https://psychologyfanatic.com/feedback-loops/

NATO Science & Technology Organization. (2024). *The understanding of cognitive warfare in comparative perspective*. NATO STO Meeting Proceedings. https://publications.sto.nato.int/publications/STO%20Meeting%20Proceedings/STO-MP-HFM-361/MP-HFM-361-P13.pdf

Nestor, M.W. and Wilson, R.L. (2022). *Anticipatory ethics and the use of CRISPR in humans*. Springer.

Nissenbaum, H. (2009). *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Platt, M., Platt, D., & McBurney, P. (2024). Sybil attack vulnerability trilemma. *International Journal of Parallel, Emergent and Distributed Systems*, *39*(3), 446–460. https://doi.org/10.1080/17445760.2024.2352740

Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press.

Singer, P. W., & Brooking, E. T. (2018). *LikeWar: The Weaponization of Social Media*. Eamon Dolan/Houghton Mifflin Harcourt.

Sociotechnical Imperative. Sociotechnical Imperative → Area → Resource 5. Accessed 2/5/2026.

Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Tavani, H.T. (2009). *Ethics and Technology: Ethical Issues in an Age of Information and Communication Technology*. Wiley.

Tomassi, A., Falegnami, A., & Romano, E. (2024). Mapping automatic social media information disorder: The role of bots and AI in spreading misleading information in society. *PLOS ONE*, *19*(5), e0303183. https://doi.org/10.1371/journal.pone.0303183

US Cyber Command. (2024, September 3). *Russian Disinformation Campaign "DoppelGänger" Unmasked*. https://www.cybercom.mil/Media/News/Article/3895345/russian-disinformation-campaign-doppelgnger-unmasked-a-web-of-deception/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380), 1146-1151.

Wilson, Richard. Anticipatory Ethics as a Method for Teaching Engineering Ethics. 2021 ASEE St. Lawrence Section Conference.

Wilson, Richard L. Cambridge Analytica, Facebook, and Influence Operations: A Case Study and Anticipatory Ethical Analysis, Proceedings of the 18th European Conference on Cyber Warfare and Security, University of Coimbra, Portugal, Academic Conference and Publishing International, Ltd. 2019.

Wilson, Richard L., Cyber Warfare, Terrorist Narratives and Counter Terrorist Narratives: An Anticipatory Ethical Analysis, Proceedings of the 18th European Conference on Cyber Warfare and Security, University of Coimbra, Portugal, Academic Conference and Publishing International, Ltd. 2019.

Wilson, Richard L. Information Warfare: Fabrication, Distortion and Disinformation: A Case Study and Anticipatory Ethical Analysis, Proceedings of the 18th European Conference on Cyber Warfare and Security, University of Coimbra, Portugal, Academic Conference and Publishing International, Ltd. 2019.

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.

Cyber Operations and the Threshold for Cyber Warfare: Ethical and Anticipated Ethical Issues

Richard Wilson[1,2] and Noah Donnelly[2]
[1]Department of Philosophy, Towson University, Baltimore, Maryland, USA
[2]Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA

wilson@towson.edu
ndonnel1@students.towson.edu

Abstract: Determining whether a cyber operation meets the threshold for being designated cyber warfare involves ethical, technical, and strategic criteria. These are primarily derived from international law frameworks such as the Tallinn Manual 2.0 and principles of the UN Charter. What are the primary criteria for cyber operations to be called acts of Cyber Warfare Threshold? (1). Scale and Effects of cyber operations. The cyber operation must cause effects comparable to a kinetic attack, such as: Physical destruction (e.g., damaging power plants, pipelines). Loss of life or serious injury. Severe disruption of essential services (power, water, healthcare). Example: Stuxnet physically damaged Iranian centrifuges, meeting this criterion. (2). The Intent and Purpose of the Cyber Operation. The cyber-attack must be strategically motivated and aimed at: Weakening an adversary's military or economic capability, Coercing or punish a state. Achieving political or military objectives. Example: Ukraine power grid attacks were linked to geopolitical conflict. (3) Cyber operation must be attributed to a State. The cyber operation must be attributable to a state or conducted under its direction/control. State sponsorship or direct involvement elevates an attack from the level of being a cybercrime to the level of acts of cyber warfare. Example: WannaCry and NotPetya were attributed to North Korea and Russia respectively. (4) There must be severity and consequences related to cyber operations. The impact of cyber operation must be significant enough to trigger international legal obligations, such as a Breach of sovereignty. The possible classification of the cyber operation as a "use of force" under Article 2(4) of the UN Charter. Example: Triton/Trisis targeted safety systems in petrochemical plants, risking catastrophic damage. (5) The Target Type of cyber operation. Attacks by the cyber operation on critical infrastructure or military systems are more likely to meet the threshold. Civilian systems may also qualify if disruption is widespread and severe. (6) The Context and Scale of the cyber operation. The cumulative effects of cyber operations mean that: repeated or coordinated attacks may collectively meet the threshold even if individual cyber operations do not. This analysis will identify the ethical and anticipated ethical issues with the identification of Cyber Operations and the ethical criteria for these operations to reach the threshold for being classified as acts of Cyber Warfare. Our methodology will employ an interdisciplinary method that draws upon distinctions taken from Bratman's BDI Model of rational agency, computer science, and conceptual ethical analysis with reference to specific case studies.

Keywords: Cyber warfare, Tallinn manual 2.0, Article 2(4), Stuxnet, Attribution, Kinetic equivalence, Critical infrastructure, Gray zone conflict

1. Introduction

In classic Westphalian systems of international relations, what acts of war were often historically ambiguous. War was defined by physical acts with immediate and catastrophic consequences. Such actions could be armies crossing sovereign borders, a missile striking a city, or naval blockades. The physical violation of sovereign territory was the primary mode of starting a conflict. However, with the advent of the digital domain, what has now been introduced is a gray zone of conflict (Clement, 2018). The gray zone is a spectrum of hostility that exists below the threshold for armed conflict, but far above peace. This is where Cyber Operations lay, where states now regularly perform digital espionage, sabotage, subversion and coercion that cause quantifiable damage to their opponent without ever meeting the legal definitions of being called a war. The defining question facing the international community in the digital age is defining the exact moment at which a line of code is no longer a tool of intelligence, but an act of war (Kello, 2013). This paper argues for a fundamental paradigm shift in how international law evaluates cyber conflict, aiming to resolve current ambiguity. The current "Effects-Based" model waits for a cyber-attack to cause physical destruction equivalent to a kinetic weapon. We contend this approach is dangerously reactive and insufficient for the digital age. Instead, we propose a "Norms-Based" threshold. Under this framework, the act of war is defined not by the volume of damage caused, but by the nature of the target itself (e.g., hospitals, nuclear plants), regardless of whether the attack succeeds or fails.

The "Threshold Problem" is not just about semantics or academic debates, defining it is imperative to guaranteeing global stability. The definition of this threshold for cyber war would decide the rules of engagement for major powers on the global stage. If a state is able to define cyber-attacks on its power grid as an act of war, which in turn leads to claims the right to self-defense, this could lead to kinetic retaliation with conventional weapons under Article 51 of the UN Charter. This violates defense mechanisms of the escalation ladder where an act of digital espionage can spark a nuclear conflict. On the other hand, if states set a threshold

for cyber warfare that is too high, like demanding a physical act, adversaries would be allowed to erode a nation's sovereignty, steal intellectual property, and cause economic harm to a targeted nation with immunity. The criteria needed to establish this threshold are Scale, Effects, Intent, Attribution, Severity, Target Type, and Context. These are all derived from the *Tallinn Manual 2.0* on the *International Law Applicable to Cyber Operations* (Schmitt, 2017). This academic study represents a comprehensive attempt to apply existing international law to the cyber domain but fails to provide the application to clear up any ambiguity or interpretations.

This ambiguity creates a problematic ethical landscape where the rules for defining cyber warfare are decided by the aggressors. A logic bomb placed in a petrochemical plant, like Triton malware (Sobczak, 2019), is meant to cause destruction of infrastructure and loss of life, but if does not detonate due to an error in the code, international law struggles to classify what kind of act has occurred. Is it an act of war or an act that attempts sabotage? Does the effect of thousands of small hacks against the financial sector equate to a single large act that would be similar to a missile strike on a bank? This analysis will investigate these questions through the lenses of Just War Theory and Anticipatory Ethics. By examining case studies, including physical destruction caused by Stuxnet, the economic chaos of NotPetya which was the targeted disruption of the Ukrainian power grid, and the lethal intent of Triton. We argue that the current ambiguity is not an anomaly, but a feature of modern conflict. It is a feature which is exploited by state actors to project power while avoiding accountability. We conclude that ethical stability necessitates a transition from "Effects Based" definitions, which are reactionary in nature, to "Norms Based" which would prohibit targeting civilian infrastructure regardless of outcome (Farrell & Glaser, 2017).

2. Technical Issues

To determine a threshold for Cyber Warfare is complicated by the unique technical architecture of the internet, it resists the precise, binary categorizations of kinetic attacks. Unlike physical weapons, which are predictable in yield and trajectory, cyber weapons are often unpredictable in strength and attribution for cyber-attacks is difficult.

The main technical metric used by international lawyers to analyze cyber warfare is Kinetic Equivalence. Within the framework of the Tallinn Manual, a cyber operation qualifies as a "Use of Force" if its scale and effects are equivalent to physical operations while also rising to a level in which use of force is justifiable. Technically, this would require measuring the impacts of digital code on physical objects, people or computer systems. One example is the Stuxnet worm causing physical destruction of Iranian nuclear centrifuges by altering the way that the Iranian nuclear centrifuges operated in order to destroy them by manipulating the SCADA (Supervisory Control and Data Acquisition) systems monitoring the centrifuges. The bridge of cyber to physical damage is the clearest technical mark of warfare because it results in broken machinery. Many cyber operations do not break physical things; they merely alter or destroy data. The technical challenge is in the quantifying of equivalence between cyber and physical acts. If a ransomware attack such as WannaCry shut down the hospitals of a nation, causing no structural damage to buildings, but causing patient deaths due to delayed care and the inability to access medical records, is the alteration of code now capable of being classified as a weapon and does this rise to the level of an act of war? The difficulty lies in the attribution of causality. In a complicated and interconnected system, proving that a specific data packet caused a specific physical outcome is a challenge that continues to evade certainty (Rid, 2012).

A prerequisite for a declaration of war is identifying the enemy, but in the cyber domain attribution is a forensic challenge shrouded in obfuscation by the use of False Flags. In the physical world, a missile has a clear return address if its trajectory is tracked. Within the cyber domain, state actors can route their attacks through proxy servers that are in neutral countries, using purchased criminal malware to mask identity, or purposely insert snippets of code in foreign nations with the intention of framing other countries and masking their own identities. For example, the malware entitled Olympic Destroyer used against the Pyeongchang Winter Olympics contained code that looked like it was written by North Korean hackers (Lazarus Group), but in fact was produced by Russia's GRU which was discovered through forensic analysis (Greenberg, 2018). Technically, establishing the responsible party with a degree of certainty that would justify a kinetic military strike would take months to years of signals intelligence analysis. The "Time to Attribution" lag renders immediate "Right to Self Defense" inoperable in real-time. A nation cannot launch a retaliatory strike if eight months are required to determine who fired the 1st cyber shot.

In addition, state actors often employ tactics such as small-scale intrusions, that remain below the threshold of war to circumvent thresholds, but that collectively equate to a strategic assault. Detecting these activities

requires advanced Correlation Analysis over a wide array of disparate datasets. Scanning a port server in Ohio, phishing attacks sent to a Texas contractor, and a server outage in New York may appear unrelated, but in fact may not be. To pinpoint the link between these attacks would take a centralized cyber defense agency, recognizing that these events represent a coordinated campaign by a single Advanced Persistent Threat (APT). The inability of dispersed civilian networks to aggregate and correlate this data in a timely manner, prevents victim states from knowing they are at war until the campaign has caused maximum damage, and the perpetrators have disappeared. This tactic exploits the time lag of bureaucratic decision making, which offers little defense against the light speed efficiency of fiber optic networks (Rid & Buchanan, 2015).

3. Ethical Issues

Norms are being degraded because of the vacuum caused by the lack of clear definition for a threshold for cyber-attacks escalating to the level of cyber warfare. The application of Just War Theory to cyberspace to reveals inconsistencies between traditional military ethics and the realities of the cyberspace in which modern cyberwarfare is conducted.

The first ethical conflict arises from Sovereignty and Article 2(4) of the UN Charter. Article 2(4) prohibits the "threat or use of force against the territorial integrity or political independence of any state." The problem is that cyber operations often violate sovereignty without the use of physical force. The ethical issue lies in the notion of Westphalian Sovereignty being eroded. Russian interference in the 2016 U.S. election did not destroy voting machines, it violated the political independence of the United States in a significant way. If the threshold for war is based strictly on physical destruction or Kinetic Equivalence, it legitimizes non-destructive operations that become existential threats to a nation's survival rather than kinetic warfare and border disputes. This essentially provides a defense for psychological and political warfare, allowing states to destabilize rivals without any fears of kinetic military repercussions (Lucas, 2017).

The second ethical issue is Retaliation as defined in Article 51 of the UN Charter. Article 51 permits self defense in case of an armed attack. If a cyber operation crosses the threshold for this, the victim state is permitted ethically and legally to respond with a kinetic attack (Waxman, 2013). However, the attribution problem creates an ethical risk, Mistaken Retaliation. If state A hacks state B, but routes attacks through an innocent state C, and state B then attacks state C, innocents suffer due to errors in forensic analysis. Employing utilitarian ethics, the escalation from cyber conflict to nuclear or conventional attacks would suggest the threshold for attribution must be absolute. Demanding 100% certainty in attribution for cyber-attacks emboldens aggressors because they can now hide within the margin of doubt. This creates an ethical dilemma between defending the citizens of one's nation, and the duty to avoid starting unjust wars due to faulty intelligence.

Finally, the Principle of Distinction would need belligerents to distinguish between military and civilian targets. In cyberspace, this is basically impossible due to the fact that the military uses civilian internet for its operations. The backbone of the internet are fiber optic cables, routers, and DNS servers which are Dual-Use Infrastructure. Attacks intended to degrade an enemy militaries command-and-control network will in turn deny civilian internet access and affect banking and emergency services. Almost any strategic cyber warfare operation carries high risks of violating Jus in Bello (conduct in war). Accepting collateral damage in cyberspace has become normalized to a degree that is incredibly dangerous. States like Russia, as seen in the NotPetya attack, treated the global economy as a "free fire" zone. The ethical failure here lies in the refusal of states to separate military and civilian networks, which holds the civilian population hostage to military operations.

4. Case Studies

Boundaries for the Cyber Warfare Threshold have not been drawn by treaties, they have been drawn by precedents that occur in conflicts. These cases show how the threshold is tested, perverted, and defined by state actors which has led to arbitrary definitions.

4.1 Stuxnet

The Stuxnet worm which was deployed in 2010 and is attributed to a joint U.S.-Israeli operation, shows a paradigmatic example of a cyber operation that met the threshold of an armed attack. Stuxnet was designed with one purpose: the physical destruction of Iran's nuclear enrichment capabilities at Natanz (Zetter, 2014). The malware targeted Siemens Step7 software used to control Programmable Logic Controllers. It modified code sent to the frequency converter drives, which caused centrifuges to spin at supersonic speeds until they broke from mechanical stress. Simultaneously, the malware performed an attack on the monitoring systems by feeding

prerecorded "normal" data to the control room so the operators would not be aware of the destruction of the centrifuges on the plant floor.

A Threshold Analysis confirms that Stuxnet met the criteria of Scale and Effects for cyberwarfare by causing verifiable physical destruction. The "Target Type" criteria were met by attacking military infrastructure. While the Stuxnet attack met the threshold of an armed attack, the muted response of the international community represented a tacit acceptance of cyber sabotage, especially in regard to nuclear proliferation. Stuxnet presents a Utilitarian outcomes-based argument: the outcome of the attack delayed nuclear proliferation without causing human loss of life, framing it as a sanitary alternative to a kinetic airstrike. A Deontological duty-based analysis would argue that the criteria of "Scale and Effects" for cyber warfare was met by the undeniable physical destruction of the centrifuges, opening a "Pandora's Box," by validating the use of malware for physical destruction (Zetter, 2014).

4.2 Ukraine Power Grid Attacks

In 2015 and 2016 the Sandworm team, a group of Russian hackers compromised the SCADA systems of Ukrainian power distribution companies, leaving hundreds of thousands of people without electricity during the peak of winter (Greenberg, 2019). The attacks started with spear-phishing emails which contained "BlackEnergy 3" malware. The malware granted access to corporate networks from which the attackers pivoted to SCADA networks, acquired VPN credentials, and remotely accessed the control panels. The operators of power companies watched helplessly as their cursors moved across their screens, opening breakers and cutting the power to substations, all in a coordinated strike. Without a means to prevent the breakers from being switched back on, the attackers had to launch a "KillDisk" wiper component that erased the boot records of the operator stations, which would prevent any attempts to restore power remotely. On top of wiping the operator stations, they flooded the customer support center with Telephony Denial of Service (TDoS) attacks to prevent customers from reporting outages.

Threshold Analysis revealed that operation Sandworm was testing the "Severity and Consequences" criteria of cyber warfare. By causing no physical destruction or direct loss of life, NATO did not invoke Article 5 and operation Sandworm hovered firmly within the gray zone. By stopping short of permanent kinetic damage, the aggressor avoided triggering a war while causing adverse effects for the targeted population. The attack violated the ethical norm which protects civilian infrastructure, in this case power stations during peacetime, effectively treating the Ukrainian populus as a test subject for cyber weaponry (Greenberg, 2019).

4.3 NotPetya

In 2017, the Russian military intelligence (GRU) unleashed the NotPetya malware, disguising it as a legitimate tax software update for the Ukrainian accounting program. Unlike standard ransomware that encrypts a company's data for profit, NotPetya was designed to wipe data, an action that is purely destructive in nature. The malware utilized "EternalBlue" which is an exploit stolen from the U.S. National Security Agency (NSA), designed to spread rapidly through the Server Message Block (SMB) protocol of the Ukrainian accounting program. The original target was Ukraine, but the autonomous nature of the malware caused it to spread globally, infecting corporations like Maersk, FedEx, and Merck. The attacks costed an estimated $10 billion in damages, crippling global shipping and logistics for weeks (Greenberg, 2019).

Threshold Analysis of NotPetya obscures conceptual distinctions related to "Target Type" and "Intent." The intended target was Ukraine, but it became a global weapon of mass destruction which was due to its wormlike nature. The scale of the financial damage caused by NotPetya was equivalent to a kinetic war, yet every state ended by posturing, western nations publicly condemned the attack but stopped short of any meaningful retaliation. NotPetya is an example of the lack of accountability for intentionally and Indiscriminately deploying Cyber Weapons. The "spillover" effect violated the sovereignty of numerous neutral nations who were not party to the Russia-Ukraine conflict. The international community failed to classify this as an "Armed Attack" which indicates that the current threshold for a cyber-attack rising to the level of cyber warfare is too high, prioritizing the status quo over the economic health of the deeply connected global system.

4.4 Triton/Trisis

The Triton malware was a sophisticated cyber weapon built to target Safety Instrumented Systems discovered in a Saudi petrochemical plant in 2017. The unique danger of this malware lays in the targeting of the Schneider Electric Triconex Safety Instrumented System (SIS). The SIS is the last line of defense, intended to shut down a plant in case of dangerous pressure or temperature spikes to prevent an explosion within the plant. The malware

was intended to reprogram the SIS controllers creating unsafe conditions, effectively cutting the safety net protecting the plant (Johnson et al., 2017). The attack failed, but only because of a coding error in the malware which caused the safety controllers to enter into the failsafe state, shutting down the plant and alerting plant operators of the malware attack.

Threshold Analysis shows that the Triton malware showed an undeniable "Intent to Cause Loss of Life." Targeting failsafe systems is pointless if you are conducting espionage or intelligence work, its only purpose is to cause death and destruction. If not for an error, Triton could have caused a massive explosion and a release of toxic hydrogen sulfide gas which would kill workers and civilians in the surrounding area. This clearly crossed an ethical "Red Line" in civilized conflict. The coding error does not prevent the perpetrators from being blamed for the attack because their intent remains. This is a clear attempt at mass homicide via digital code, and represents a possible war crime (Sobczak, 2019).

4.5 SolarWinds

In late 2020, the "SolarWinds" supply chain compromise showed how fragile the threshold between espionage and an armed attack. SolarWinds was attributed to the Russian SVR, a foreign intelligence service, and was an operation that relied on stealthy infiltration. The attackers injected a backdoor known as "Sunburst" into the process of building software for the SolarWinds Orion network monitoring platform. The "Sunburst" code was digitally signed by a legitimate certificate which allowed it to bypass the standard trust verification in over 18,000 government and corporate networks. Once the attackers had established the backdoor, they used "Teardrop" which is a malware that moves laterally and exfiltrates data that is oftentimes sensitive. The unique challenge presented by SolarWinds is one of "Scale and Effects." The scale was massive, it compromised the Pentagon, DHS, and Treasury, the effect was not destructive in nature, no data was wiped, no systems were bricked. The Tallinn Manual 2.0 classifies this as espionage, which is not prohibited under international law (Schmitt, 2021). However, the ethical line is blurred due to the operation implanting malicious software into critical infrastructure, the physical equivalent of this would be preparing a battlefield. This blurs the line by questioning whether the act of war is merely the detonation of a bomb, or if planting explosives counts as an act of war. The SolarWinds case demonstrates the failure of current threshold analyses to take "Strategic Latency" into account, which is the capability to destroy being establish, but not immediately used (Microsoft, 2020).

5. Anticipatory Ethics

Failing to prevent these escalations in cyber attacks is a failure in foresight. Applying "Moral Responsibility for Computing Artifacts" rules by Miller et al. (2011) to design cyber weapons and targeting of critical infrastructure shows significant ethical lapses.

5.1 Rule 1: The Foreseeability of Effect

Rule 1 declares that designers are morally responsible for the foreseeable effects of their artifacts. This applies directly to state agencies developing "Zero-Day" exploits. While designing "EternalBlue" the NSA foresaw that if such a powerful tool were to be leaked, it could allow for bad actors to cause massive amounts of damage. When the Shadow Brokers leaked it, and when it was subsequently being used in the WannaCry and NotPetya attacks, the "foreseeable effect" later materialized in the form of billions of dollars in damages and risks to patient wellbeing in hospitals. Through the lens of Anticipatory Ethics, the idea is that as technology develops ethical problems will also develop, so stockpiling "Zero-Day" vulnerabilities for cyber-attacks is unethical at its core, because the state of the technology at the present time is that the weapon cannot be controlled. A nation state has an ethical duty to disclose any vulnerabilities to patch civilian infrastructure, this is a process called Vulnerability Equities Process, as opposed to hoarding them for future warfare, do not make them at all. The prioritization of offense over defense is a violation of the social contract between a state and its people (Brey, 2012).

5.2 Rule 4: The Sociotechnical Imperative

Rule 4 states that the design of computing artifacts must account for the sociotechnical systems in which they are embedded. Industrial Control Systems (ICS) are old, designed in the 1970s and 80s for a sociotechnical system where operators were inherently trusted (Knowles et al., 2015). They were not designed nor were they ever intended to connect to the modern internet. When connecting legacy systems to the web, one violates the Sociotechnical Imperative by embedding fragile systems in a hostile environment, like how in the Middle Ages forts would crumble under modern artillery fire. Therefore, Anticipatory ethicists can argue that critical infrastructure, specifically in power, water, and nuclear sectors must anticipate cyber warfare, designing systems

that have mechanical overrides and backups that run on a closed system. The reliance on insecure-by-design systems in important infrastructure is a failure of due diligence and a neglect of the moral duty to protect public safety (Miller et al., 2011).

6. Recommendations

Before outlining a normative framework, we must acknowledge the significant political and technical constraints hindering these solutions. State actors, particularly major cyber powers, benefit from the current ambiguity of the "Gray Zone." They have little incentive to sign treaties that would limit their strategic options. Technical challenges with attribution also make enforcing these norms difficult in real-time. The following recommendations offer an aspirational roadmap, a phased evolution of international law that begins with voluntary bilateral agreements before expanding into enforceable global treaties. Ambiguity must be resolved if we desire to prevent catastrophic escalation from cyber-attacks to cyber warfare to kinetic warfare. We propose a framework of Normative Constraints and Technical Defense that moves past the reactive model that is currently used.

The first recommendation is the creation of a Digital Geneva Convention (Microsoft, 2017). The international community needs to establish a treaty that clearly prohibits attacks targeting civilian infrastructure, regardless of the “Kinetic Equivalence.” Attacks targeting civilian infrastructure should be disallowed regardless of them being digital or physical. Attacks on the public sectors of the internet (like DNS root servers and BGP routing tables), healthcare systems, and nuclear safety systems must be designated as involving clear acts of war. This disposes of the ambiguous “effects-based” analysis. Under an effects-based analysis, if an attack on a hospital killed zero patients, it would be ignored, under a norms-based framework the act of attempting to target the hospital would be a war crime alone. This restores the Principle of Distinction by creating recognized safe zones in cyberspace that are off-limits to cyber-attacks and cyber conflicts (Finnemore & Hollis, 2016).

The second recommendation is implementing Mandatory Attribution Markers, acting like Shoulder Insignia. States must adopt standards where cyber operations performed by state actors require a cryptographic signature which identifies the state of origin. While this is counter intuitive to espionage, it follows the Geneva Convention standard of combatants wearing uniforms and carrying arms openly. If a state desires the legal protections of “Combatant Immunity” for its hackers, the hackers need to identify their country of origin or be tried as criminals in the affected state. This reduces the risk of misidentifying the aggressor, and in turn the additional problem of unidentifiable hackers striking the wrong target.

The third recommendation is to create Active Defense strategies. States must transition from passive defense (patching and firewalls) to active defense (hunting state sponsored hackers) (Denning, 2014). The U.S. Cyber Command strategy of ‘defending forward’ employs the tactic of disrupting attacks at the source before they can cause damage. In the realm of ethics, this lowers the risk of damage by neutralizing threats early, but must in turn be heavily regulated to prevent a “Preemptive War.” Active defense should be limited to disrupting the tool or infrastructure of the attacker and must not involve killing or harming the personnel of the adversaries’ host nation.

7. Future Work

Future research must address the role of Artificial Intelligence in reacting to warfare thresholds. Cyber operations occur at machine speed; humans are too slow to keep up.

Research is required into the ethics and issues related to Autonomous Responses to cyber-attacks. If an AI defense system catches an intrusion into a power grid and the signature matches that of a known destructive wiper, should it be allowed to counter-strike without human authorization? We anticipate a future where AI driven malware and AI driven defense will fight against one another, eventually escalating conflict to a level in which humans can no longer intervene. Future work needs to focus on designing physical failsafe’s for automated cyber conflicts that automatically pause hostilities when certain markers of damage or speed are passed (Taddeo & Floridi, 2018).

8. Conclusion

The threshold for determining when and how cyber-attacks rise to the level of cyber warfare persists as a dangerous ambiguity in geopolitics. The reliance on “Kinetic Equivalence” which is the idea that a cyber attack needs to look like an explosion to count at war is incredibly outdated. The case studies of Stuxnet, NotPetya, and Triton show that state actors are persistently towing the line. They exploit the Gray Zone to undermine

sovereignty and project power without having to declare war and face the consequences any nation conducting warfare would typically endure.

By applying Anticipatory Ethics, we come to the conclusion that the "Effects-Based" model is not sufficient for defining the threshold for defining when a cyber-attack rises to the level of cyber warfare because it relies on a post-hoc analysis. Waiting for damage to occur before calling something a cyber-attack or an act of cyberwarfare is a bad method of operating. To ensure a prosperous digital future, we must move towards a Norms-Based model which fundamentally moves the threshold for cyber warfare from consequentialism (damage-based judgement) to a deontological framework (judging an attack based on what its target is). The dissimilarity from the "Effects-Based" model, which requires a victim state to wait for kinetic equivalence, like physical destruction or the loss of life before any operation is considered an act of war. A Norms-Based model would create proactive "Red Lines" concerning categories of protected infrastructure. Within this framework the act of targeting systems that are indispensable to civilian survival, like hospital grids or election infrastructure, now becomes a violation of international norms and is classified as an act of aggression regardless of their success or lack thereof. This approach removes any ambiguity that the "Gray Zone" granted previous operations by focusing on intent as opposed to physical impact. Without a clear and enforceable definition of when a cyber-attack becomes an act of war and for what constitutes war in the cyber domain, we cannot hope for stable peace in the future.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.1 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing.

References

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. https://www.acm.org/code-of-ethics

Bratman, Michael. (1999). Intentions, Plans and Practical Reason. Center for the Study of Language and Information.

Brey, P. A. E. (2012). Anticipatory Ethics for Emerging Technologies. *NanoEthics*, 6(1), 1–13.

Clement, J. A. (2018). Cyber Warfare and the Gray Zone. *The Cyber Defense Review*, 3(2), 53–62.

Denning, D. E. (2014). The Ethics of Cyber Conflict. *The Handbook of Information and Computer Ethics*, 407–426. Wiley.

Greenberg, A. (2019). *Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers*. Doubleday.

Greenberg, A. (2018). The Untold Story of the 2018 Olympics Cyberattack, the Most Deceptive Hack in History. *Wired*. https://www.wired.com/story/untold-story-2018-olympics-destroyer-cyberattack/

Johnson, B., Caban, D., Krotofil, M., Scali, D., Brubaker, N., & Glyer, C. (2017). *Attackers Deploy New ICS Attack Framework "TRITON" and Cause Operational Disruption to Critical Infrastructure*. FireEye Mandiant.

Kello, L. (2013). The Meaning of the Cyber Revolution: Perils to Theory and Statecraft. *International Security*, 38(2), 7–40.

Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., & Jones, K. (2015). A Survey of Cyber Security Management in Industrial Control Systems. *International Journal of Critical Infrastructure Protection*, 9, 52–80.

Lucas, G. R. (2017). *Ethics and Cyber Warfare: The Quest for Responsible Security in the Age of Digital Warfare*. Oxford University Press.

Microsoft. (2017). *A Digital Geneva Convention to protect cyberspace*.

Microsoft. (2020). *Deep Dive into the Solorigate Second-Stage Activation: From Sunburst to Teardrop*. Microsoft Security Response Center.

Miller, K.W., et al. (2011). Moral Responsibility for Computing Artifacts: 'The Rules'. *IT Professional*, 13(3), 57–59.

Rid, T., & Buchanan, B. (2015). Attributing Cyber Attacks. *Journal of Strategic Studies*, 38(1-2), 4–37.

Rid, T. (2012). Cyber War Will Not Take Place. *Journal of Strategic Studies*, 35(1), 5–32.

Schmitt, M. N. (Ed.). (2017). *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*. Cambridge University Press.

Schmitt, M. N. (2021). SolarWinds and the International Law of Cyberspace. *The Cyber Defense Review*, 6(2), 29–46.

Sobczak, B. (2019). *The Inside Story of the World's Most Dangerous Cyberattack*. E&E News.

Taddeo, M., & Floridi, L. (2018). Regulate artificial intelligence to avert cyber arms race. *Nature*, 556(7701), 296–298.

Waxman, M. C. (2013). Cyber-Attacks and the Use of Force: Back to the Future of Article 2(4). *Yale Journal of International Law*, 36, 421–459.

Zetter, K. (2014). *Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon*. Crown.

Lessons Learned While Performing 'Hands On' Teaching in a Live Environment

Adamu Yusuf, Justin Clevenger and Char Sample
Marshall University, Huntington, WV, USA

charsample50@gmail.com

Abstract*:* In West Virginia (WV) the county sheriff's office is a central point of interaction between the local government and the people who are served by these elected officials. While law enforcement extends well beyond, this central point of contact, typically the county sheriff website, is not only a great starting point for the citizens to interact with local law enforcement, but also an entry point for potential illegal activities. The county sheriff's offices websites in WV are also the central point for online payment of local taxes, thus making these sites a potential target for electronic misdeeds that would range from defacement to theft and potentially beyond. This study team examined the websites for the 55 counties of WV, in part as a classroom exercise, for potential vulnerabilities ranging from standard server scans to potential exposure through social media (SM) sites where information can be inferred. The majority of SM usage showed professional pages on LinkedIn and Facebook and newer SM sites were not used. The study findings suggest that an overwhelming majority have outsourced services to third party cloud service providers. Overall WV sheriff offices practice good cyber hygiene however, two areas of potential problems are in the duplicate IP addresses found in the cloud service providers shared servers, running multiple services with multiple tenants on the same physical server and the existence of legacy information. The study was neither comprehensive, nor intrusive in nature, but did serve to verify the security posture of the various counties sheriff offices allowing for a fingerprinting of a group of sites in the same industry allowing for a general overall view of security by sector in the public service space. The study also suggested the need for periodic re-checking suggesting a less intense degree, a practice that should be performed regularly in any secure architecture and acts as a public record for interested citizens. A final lesson from the class experience re-iterates the importance of proper scoping.

Keywords: Scans, Tools, Social media, Cloud service providers, Third party, att4ck framework

1. Introduction

Cybersecurity research increasingly focuses on aggregate, posture-based assessments to understand systemic risks and shared vulnerabilities across public-sector organizations, rather than relying solely on incident reporting. ENISA's sectoral threat landscape for public administration highlights that government entities face common threat patterns that can be identified through large-scale, non-intrusive analysis of externally observable infrastructure (European Union Agency for Cybersecurity [ENISA], 2024). ENISA's annual Threat Landscape further emphasizes the value of cross-sector aggregate analysis for identifying emerging threats, shared attack vectors, and trends that go beyond individual organizations (ENISA, 2025). In the United States, CISA similarly promotes proactive cyber hygiene and external attack surface assessments to help state and local governments reduce exposure to vulnerabilities, especially in environments with limited resources and legacy systems (Cybersecurity and Infrastructure Security Agency [CISA], 2023).

The 2025 Marshall University Vulnerability Assessment class was abnormally small and class size afforded the group the opportunity to pick a meaningful target as long as the effort was limited to reconnaissance techniques and the targets were given informed consent, as directed by the Marshall University IRB. Based on this, the sheriff offices for the 55 counties of West Virginia were selected as the target and the letters were sent. The embedded link provides a copy of the letter l Letter Template - justin - char.docx.

2. Background

Research on local governments shows that these entities often face increased cyber risks due to fragmented governance, limited budgets, and dependence on shared infrastructure. This underscores the need for a comprehensive, posture-based evaluation (Hossain et al., 2024). Empirical evidence also suggests that cybersecurity incidents in municipalities have systemic financial consequences; an analysis of over 1,000 cyberattacks against public entities found that municipal borrowing costs increase after publicly reported cyber incidents, reinforcing the broader institutional risks of inadequate cybersecurity posture (Curti et al., 2024).

A recent industry assessment ranked West Virginia among the top states in cybersecurity; however, such evaluations usually focus on reported cyber incidents and state-level malware statistics rather than on proactive assessments of potential entry points within specific government sectors. Prior research has shown that cyber intrusions are often discovered months after the initial compromise, with average detection delays approaching six months, which can obscure underlying vulnerabilities and delay remediation (McDermott, 2025; Total Assure,

2025). Therefore, posture-oriented analyses that examine observable infrastructure characteristics before an incident occurs offer a more forward-looking and insightful perspective on cybersecurity risk.

Public-facing government websites are recognized as essential parts of the public-sector attack surface. Research on government web infrastructure consistently reveals issues like outdated software, exposed services, and weak configurations, which increase the risks of defacement, data theft, and service disruptions (European Union Agency for Cybersecurity [ENISA], 2022). In law enforcement agencies, these risks are even higher because websites act as the main contact points with citizens, including access to public records and online payment systems.

Beyond technical exposure, prior research has emphasized the broader social effects of cybersecurity failures in government systems. Shandler and Gomez (2022), using data from multiple government websites, examined the psychological and social repercussions of cyberattacks on public institutions and found that such incidents can significantly diminish public trust, causing anger, uncertainty, and decreased confidence in affected agencies. While their work focused on results after incidents, it highlights the importance of analyzing government systems as a whole to understand the potential impacts of widespread vulnerabilities.

Several studies have used aggregate methods to analyze government digital environments beyond just incident response. Neumann et al. (2022) examined hundreds of government websites to identify patterns in political communication and public engagement, demonstrating the value of large-scale evaluations of public-sector web presence. The increasing reliance of government agencies on outsourced and cloud-based services highlights the necessity for comprehensive cybersecurity assessments. Sector-level studies have found that shared hosting environments and third-party service providers can pose systemic risks related to multi-tenancy, duplicate IP addresses, and dependence on shared infrastructure components (Bayuk et al., 2018; ENISA, 2022). These risks are especially critical for county-level agencies, where multiple jurisdictions may depend on the same providers, even though they operate independently.

Alongside traditional web and hosting infrastructure, social media platforms have become vital communication tools for law enforcement and other public agencies. While these platforms improve public engagement and facilitate information sharing, overall evaluations of public-sector social media use indicate that these tools are often managed outside formal IT governance frameworks, increasing the risk of data leaks, account breaches, and reputational harm (Cybersecurity and Infrastructure Security Agency [CISA], 2023). Oversight-focused reviews of public organizations also reveal that gaps in governance, risk assessment, and security controls are common across government settings, even when handling sensitive information systems (U.S. Government Accountability Office [GAO], 2017). Despite their operational significance, social media accounts and other visible external assets are frequently excluded from formal cybersecurity posture evaluations.

Although existing literature demonstrates the value of aggregate cybersecurity evaluations in public-sector contexts, limited research has focused specifically on county-level law enforcement agencies or combined analysis of web services, shared hosting dependencies, and social media presence within a single study. This gap motivates the present work, which applies a non-intrusive, properly scoped, aggregate assessment to county sheriff offices in West Virginia. By examining publicly observable infrastructure across a defined sector, this study contributes to a broader understanding of shared cybersecurity practices, common exposure patterns, and overall cybersecurity hygiene in local government environments.

Rather than focus on the aftereffects of failed cybersecurity (Shandler & Gomez, 2022), this study chose a neutral stance on outcomes and was observational in nature. However, findings that would indicate widespread vulnerabilities could predict future breakdowns in the trust between law enforcement and the citizenry or could become a re-enforcement of trust between law enforcement and the population. While this study was not as ambitious as the Shandler & Gomez (2022) study, the cyber aspect with aggregated government data does suggest that this study is of potential interest beyond the confines of West Virginia. The choice of the sheriffs' offices allowed the team to examine a real world, live set of targets while making those targets relatable to the students. The traditional controlled method of using a known site with known vulnerabilities as are used in various cyber exercises provides value to students but is not an accurate reflection of reality and may not provide the best opportunity for students to refine report writing when findings are minimal or even artificially high.

The research question posed was: Using basic reconnaissance techniques, what can be learned about the West Virginia sheriff sites collectively, Or, in other words, is there a collective story to be told about the 55 counties sheriff offices.

3. Method

The data collection phase required a manual search of all 55 counties of WV to determine if the sheriff's office information was available online. All 55 counties had websites that were active. A list was compiled and each site was tested to verify the site was active. Data collection on site data is described later in this section.

The data processing relied on a combination of tools and manual queries. The team was restricted to passive information gathering and the MITRE ATT4CK framework (attack.mitre.org, 2026) enumerate various techniques that are part of the reconnaissance effort. There are 11 techniques listed as part of the reconnaissance phase in assessing vulnerabilities, the team was successful in 5 of the 11 reconnaissance techniques. Table 1: Lists the techniques and provides a status on the attempt and results.

Table 1: MITRE ATT4CK enumeration of techniques and results

Technique	Results & Tool Deployed
Active Scanning	Success, Nmap
Gather victim host information	Success Nmap
Gather victim identity information	Success, Manual searches of social media sites
Gather victim network information	Not successful, no access given and would have required active measures
Gather victim org information	Success, manually accomplished via review of public information
Phishing for information	Not successful, prohibited
Search closed sources	Not successful, limited resources
Search technical databases	Not attempted
Search open websites and domains	Success, domain info via dig, nslookup, other info via Nmap
Search Threat Vendor Data	Not attempted, not available
Search victim owned websites	Not attempted, not available

The team first tested the connectivity of all of the sites ensuring they were up and active. A simple ICMP echo request was performed and a positive response from the site was noted. The lack of a positive response required a traditional web request as some sites block ICMP requests. ICMP requests are faster to evaluate, thus ICMP requests were the preferred initial method of determining connectivity. The **Nmap** scanning tool, while user friendly is time consuming and provides additional information that required additional filtering.

Upon completion of site identification, site gathering domain information followed. The domain information yielded results that gave additional information about the site and what services are outsourced and to whom. The class then invoked the **dig** (Liu et al., 2006) and **nslookup** (Kavieskara, 2021). tools. Students were specifically directed to dig for MX (mail exchange), CNAME (canonical name) for web and other servers, NS (name server) and SOA (start of authority) records in addition to the A records. These records were specifically chosen because they are minimally required for an internet presence and they can provide insight into trusted third parties.

DNS lookups using the **dig** command or the **nslookup** program were performed to gain additional knowledge about the zones. Rather than exploring a group of IP addresses the team examined host names to better understand the nature of the relationships between public hosts in the same domain. Accomplishing this task required more detailed lookups of each of the domains, focusing on specific CNAME, A, MX, NS and SOA records.

The SOA, NS and MX records provided information on the cloud service providers that had been sourced to perform the name service and mail processing. The NS records provided the list of authoritative nameservers for the zone, along with the information on who administers the zone. The MX records point to the host that accepts mail for the zone, once a MX record is retrieved the corresponding A record can give information about the software that accepts the zone's mail.

By gathering the additional information of outsourced sites, the Nmap scans were directed beyond the IP range associated with the webservice and allowed for collection of information about various cloud service providers. While individual DNS queries were time consuming, the information obtained provided insights into sheriff offices and trusted third parties.

All requests and responses were entered into a shared spreadsheet, The initial list of the 55 counties can be found at this link, West Virginia Counties and the shared spreadsheet can be found at Marshall-615.

After the DNS information was captured, port scans were performed on the hosts, and IP address blocks. That results from that scans were put in individual files, and the spreadsheet had the list of IP addresses and open ports for each county. The addresses were sorted and grouped with the intent of finding possible shared spaces, open or unsecured ports and other information.

The final reconnaissance effort was dedicated to social media findings. The social media sites that were probed were Facebook, LinkedIn, Twitter (X) and BlueSky. A test was run comparing the output of popular AI algorithms against the results of manual queries to each of the social media sites. Pilot, ChatGPT, Gemini and Grok were each tried and all of these AI engines were not as accurate as the manual queries.

All results were validated first by the instructor. Nmap scans were run multiple times on targeted IP addresses. Additionally, Nmap queries on ports were manually verified via directly by connecting to the port comparing responses. DNS queries were cross checked with **dig** commands being run on **nslookup** outputs, and **nslookup** commands were run against **dig** commands in each case comparing the outputs.

4. Findings

There were 2 common configurations identified on the web servers, the first was some servers ran on port 80 (HTTP) and performed a hand-off to port 443 (HTTPS), while other servers run strictly on port 443. Furthermore, Furthermore, the HTTPS TLS version number was checked verifying the site is running version is HTTPS v1.2 or v1.3. Both versions are currently acceptable standards (Kohlweiss et al., 2015), although v1.3 is stronger and features forward secrecy.

The social media findings relied on querying the office entity, followed by the sheriff and when available deputy names. Overall, the majority of sites were deemed “safe”, a full 69% (38/55) outsourced the website management to a third-party provider. Outsourcing web and other services to third party providers is quite common and providers are presumed to have financial incentives to enact good cybersecurity practices although evidence of this presumption is lacking. Security varies based on terms of the service level agreements (Nicolazzo et al., 2024). This team did not have access to the service level agreements. What the team did see was some variations in the open ports, although overall the profile was rather consistent across sites. The most consistent configuration showed ports 80 (HTTP) and 443 (HTTPS) as the only open ports on the webserver.

While no specific exploitable vulnerabilities were identified, there were several security irregularities observed. One of the more interesting findings was the 9 cases of shared physical servers (same IP address) with 2 different counties. The cloud service providers business model relies on shared resources, but the exact same IP address for tax payment for WV counties. The team recognized the model of cloud service providers requires the sharing of devices as the cloud model relies on efficiencies where physical devices shared by the customers.

One finding of interest was the failure to fully purge previously owned domain name space. There was a county that when scanned using **Nmap** appeared to have outsourced the DNS service management to a cloud service provider (CSP), however, the **dig** and **nslookup** queries showed a legacy address with the county name but the current address is now affiliated with a business not a sheriff’s office.

Mail Servers The most common e-mail server used was Microsoft Outlook in the Microsoft Cloud (26%). The second most was a cloud service mail solution dedicated for government use (15.2%). Cloud Service Providers (CSPs) collect analytics, in the case of this second group of users, the collected analytics could be grouped with other government entities to reveal potential patterns by targeted sector.

One of the most significant threats to cloud computing is misconfigurations (CSA Working Group, 2024). 9 of the 55 counties shared a mail server with the at least one other West Virginia county. This use of CSPs also suggests a willingness of the counties to employ personnel experienced in core government functions, further investigation could confirm this suggestion.

9/55 counties (16.3%) used the same physical mail server, although not all on the same server. The greatest risk in the cloud environment occurs when cloud tenants self-manage and fail to remain current (Khalil, 2025). Container breakouts are serious, but public facing containers, even when broken are secured to avoid movement into other containers, in part due to isolation of kernel functions (Wu et al., 2020). The top 3 CSP mail solutions were Microsoft Outlook, emailsvr.com and nicusa-gl.com, these 3 sites account for 26 of the 45 active mail sites.

The CSP mail servers ran mail on port 25, showing the port "open" not "filtered". One CSP, not listed above, had additional ports open but filtered, those ports were echo (7), chargen (19) and 22 (ssh).

Of note, 3 counties, were self-managed, these counties security profiles appeared to be slightly less robust than the managed sites, in that the open ports were not filtered. The lack of probing prohibited the group from being able to determine if the "open" ports were running the most up to date fully patched software. One of the self-managed counties showed open ports for DNS (53), HTTP (80), FTP (21) and IMAP over SSL/TLS (993). The self-managed sites did limit the number of open ports on the server.

4.1 Webservers

42/55 (76.36%) counties used HTTP protocols for their tax websites. Of the counties that used HTTPS five used TLSv1.2 & TLSv1.3, two used just TLSv1.2, and two used TLSv1.0, TLSv1.1, and TLSv1.2. No information was able to be obtained from the other four IP addresses.

32 of 55 counties (58%) used the same third-party provider for their web service. There were two primary companies who managed the majority of sites, and both had ports 80 (HTTP) and 443 (HTTPS) open. Minimally port 80 should be filtered, but preferable would be to have all port 80 traffic re-directed to port 443. Considering the site is use for paying taxes the extra privacy granted by using HTTPS instead of HTTP to HTTPS would reduce the risk profile, assuming that they server is well-maintained. One county did not use port 80 and enabled port 443 for web re-direction. Two counties used port 80 exclusively thus potentially exposing private data. The possible status cannot be made any more definite due to the non-probing nature of the study. The possibility does exist that a private session occurs before data is exchanged.

4.2 Social Media

One suggestion that deviates from historic vulnerability analysis deals with the rise of social media usage. Social media presence is considered a part of the current vulnerability assessment process to the extent that social engineering red flags that may allow for attackers to perform traditional attacks like using technical means or social methods where interaction with targets and attackers take place. (Roy et al., 2022) Less effort is spent on using social media for learn more about the target for the purpose of influence operations.

Influence operations when successful can have the target perform the attacker's goals without the attacker needing to take the controls (Tayouri, 2020). Tayouri (2020) noted that influence campaigns can operate outside of the military and political theaters. This type of attack is sometimes referred to as cognitive warfare (Bardin, 2025). The goal in cognitive warfare is to hack the target's thought process.

The process for collecting this information was time-consuming and potential connections or leads were followed (e.g. Sheriff sites that had links to other personnel such as deputies were followed and the deputies were also investigated). Leads could be considered groups associated with the social media site. The team used their personal accounts to explore group memberships on Facebook, Instagram and LinkedIn.

Before examining groups, the team first checked the advertised social media presence of the sheriffs' offices. All of the sites maintained a professional presence without revealing personal preferences. The site messages reflected professional information that aligns with the county websites.

When examined in totality 50 of the 55 counties (90.9%) had a Facebook page. LinkedIn was also a popular choice where 49/55 counties (89.09%) had profiles. The third social media platform was Instagram where 6 of 55 counties (10.90%) had Instagram pages. When determining exposure levels counties with only 1 site presence were considers low exposure, counties with 2 social media sites were considered moderately exposed and those counties with 3 or more sites were considered highly exposed. The exposure has less to do with vulnerabilities than the actual presence where data can be collected and ultimately combined with other data on the targeted person to build profiles. When considered by traditional cybersecurity standards the social media usage by the sheriffs' offices would be considered secure. However, when examined from the perspective of a data scientist where any data exposure has risk, the majority of counties have medium exposure scores. Figure 1 shows the exposure rates for the WV Sheriffs Offices.

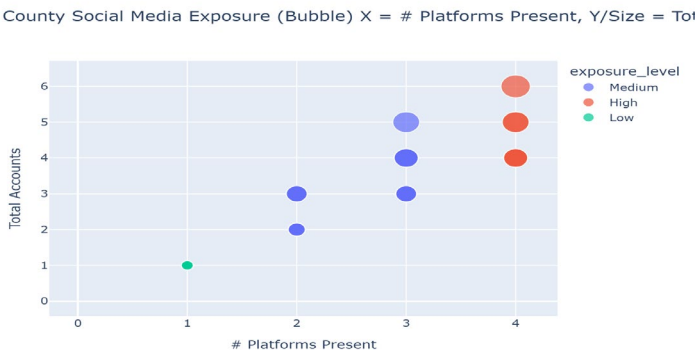


Figure 1: Exposure rates in social media for sheriffs and deputies in WV counties

Other findings: One set of findings that were of interest and ultimately prescient, were outsourcing of name service management to CSPs. Recent outages by AWS (Borngreat-Mensah, 2025) and Cloudflare (Booth, 2025) highlight the vulnerabilities to availability even when CSPs have high availability requirements. The students correctly identified a potential outage by CSP as being a potential risk area, and the recent aforementioned outages provided a real-world example showing the effects of these dependencies. Figure 2 shows the common outsourced services and the usage rate for the providers.

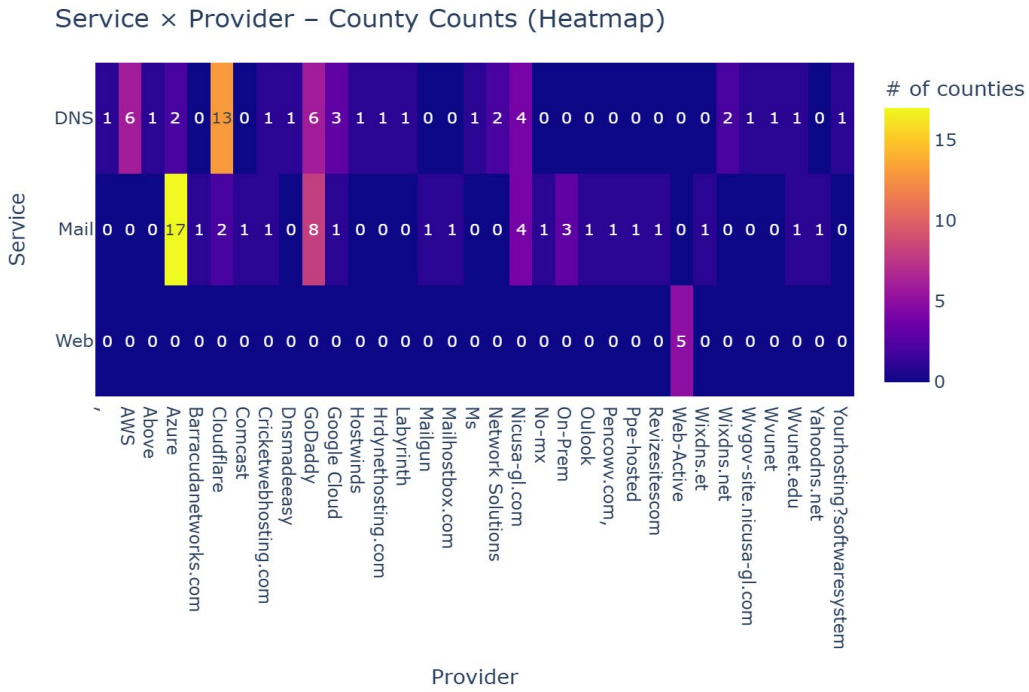


Figure 2: Providers and services, shows the rates of outsourcing critical services and most popular providers

5. Conclusions

This exercise generated a fair amount of enthusiasm at both the institution and within the class while offering a wider aggregated view of the WV Sheriffs offices. The limitations due to equipment and broad scope were difficult to reconcile. The equipment limitation resulted in Nmap scans taking longer than planned. The overall findings largely reflected industry recommendations, limited numbers of open ports and current patched software versions. The most common problems found was the HTTP to HTTPS hand-off issue.

There were individual sites, some self-managed others contractor managed that had a significant number of ports open, that raised concerns. These were not specifically listed in this paper due to privacy concerns and the fact that the county had no agreements in place with the university. These findings while uncommon, serve as outlier to the aggregated data.

The students gained knowledge of underlying technology and methods that are used by scanning tools, thereby giving the students an additional perspective. This exercise in performing non-invasive reconnaissance in a live environment allowed for deeper investigation into the sheriff sites than a basic scan while suggesting potential value in exploring inter-relationships between sites that arise when using the same vendors or cloud service providers.

Several technical recommendations were,

- The clean-up of legacy DNS information. This recommendation is made under the netiquette guidance as well as performance and security concerns.
- Failure to properly maintain or release dated DNS information can lead to slower response times (Zhang et al., 2024), or in some cases DNS hijacking (Ibid). Failure to update DNS glue records is common.
- Follow-up with individual sites that had ports not related to web services open.
- Follow-up with sites that may have potential vulnerabilities identified when compiling this paper.
- Examination of the contractor supply chains through periodic audits should be performed since most sites rely on multiple third-party vendors.

AI statement: This paper was created without the use of Artificial Intelligence (AI) when generating content. AI use was experimental in nature and described in the text of the paper.

Ethics statement: Ethical concerns were addressed at Marshall University where the institutional review board (IRB) determined that due to the non-intrusive nature of the study the IRB process was waived in favor of issuing an informed consent letter to the various counties. The informed consent letter was sent before the study started and a copy of the letter is referenced via an embedded link in the paper.

References

Bardin, J.S., 2025. Cyber warfare. In *Computer and Information Security Handbook* (pp. 1345-1380). Morgan Kaufmann.

Bayuk, J. L., Healey, J., Rohmeyer, P., Sachs, M. H., Schmidt, J., & Weiss, J. (2018). *Measuring cybersecurity risk*.

Booth, R. 2025. Cloudflare outage causes error messages across the internet. Retrieved from: https://www.theguardian.com/technology/2025/nov/18/cloudflare-outage-causes-error-messages-across-the-internet

Borngreat-Mensah, P.E.T.E.R., 2025. Mitigating cloud centralization risks: A case study of the October 20, 2025 AWS outage using a multi-cloud dependency mitigation framework (MCDMF). Retrieved from: https://www.researchgate.net/profile/Peter-Borngreat-Mensah/publication/397175468_Mitigating_Cloud_Centralization_Risks_A_Case_Study_of_the_October_2025_AWS_Outage_Using_a_Multi-Cloud_Dependency_Mitigation_Framework_MCDMF/links/6907174f4baee165918f9ccc/Mitigating-Cloud-Centralization-Risks-A-Case-Study-of-the-October-2025-AWS-Outage-Using-a-Multi-Cloud-Dependency-Mitigation-Framework-MCDMF.pdf

Curti, F., Ivanov, I., Macchiavelli, M., & Zimmermann, T. (2024). *City hall has been hacked! The financial costs of lax cybersecurity* [Conference presentation]. Municipal Finance Conference, Brookings Institution. https://www.brookings.edu/articles/what-cyberattacks-do-to-municipal-issuers-borrowing-costs/

Cybersecurity and Infrastructure Security Agency. (2023). *Cyber hygiene services*. Retrieved from:https://www.cisa.gov/cyber-hygiene-services

Cloud Security Alliance, 2024. Top threats to cloud computing. Retrieved from: https://cloudsecurityalliance.org/artifacts/top-threats-to-cloud-computing-2024#

European Union Agency for Cybersecurity. (2024). *ENISA threat landscape for the public administration sector*. https://www.enisa.europa.eu/sites/default/files/2026-01/ENISA%20Public%20Administration%20TL%202024%20-%20v1.2.pdf

European Union Agency for Cybersecurity. (2025). *ENISA threat landscape*. https://www.enisa.europa.eu/topics/cyber-threats/threat-landscape

Hossain, S. T., Yigitcanlar, T., Nguyen, K., and Xu, Y. (2024). *Local government cybersecurity landscape: A systematic review and conceptual framework*. *Applied Sciences, 14*(13), 5501. https://www.mdpi.com/2076-3417/14/13/5501

Kavisekara, H.K.M.U.I.K., 2021. *User Friendly DNS Diagnosing Dashboard* (Doctoral dissertation).

Khalil, M., *The definitive guide to cloud security compliance 2025,* DeepStrike. Retrieved from: https://deepstrike.io//blog/cloud-security-compliance-2025-guide

Kohlweiss, M., Maurer, U., Onete, C., Tackmann, B. and Venturi, D., 2015, November. (De-) constructing TLS 1.3. In *International Conference on Cryptology in India* (pp. 85-102). Cham: Springer International Publishing.

Kornfield, M., 2025. Social Security official says DOGE compromised Americans' data. *The Washington Post*, pp.NA-NA.

Liu, C. and Albitz, P., 2006. *DNS and Bind*. " O'Reilly Media, Inc.".

Madnick, S.E., Li, X. and Choucri, N., 2009. Experiences and Challenges with using CERT Data to Analyzes.

McDermott, J. (2025). Hackers linger for an average of 206 days: Here's how we can spot them on day 1, Retrieved from: https://43tc.com/article/hackers-linger-for-an-average-of-206-days-heres-how-we-can-spot-them-on-day-1/)

McLaughlin, J., (2025). A whistleblower's disclosure details how DOGE may have taken sensitive Labor data. https://www.npr.org/2025/04/15/nx-s1-5355896/doge-nlrb-elon-musk-spacex-security.

MITRE Attack website. Retrieved from http:attack.mitre.org

Neumann, M., Linder, F. and Desmarais, B., 2022. Government websites as data: a methodological pipeline with application to the websites of municipalities in the United States. *Journal of Information Technology & Politics*, *19*(4), pp.411-422.

Nicolazzo, S., Nocera, A. and Pedrycz, W., 2024. Service level agreements and security sla: A comprehensive survey. *arXiv preprint arXiv:2405.00009*.

Roy, S., Sharmin, N., Acosta, J.C., Kiekintveld, C. and Laszka, A., 2022. Survey and taxonomy of adversarial reconnaissance techniques. *ACM Computing Surveys*, *55*(6), pp.1-38.

Srikanth, B., 2022. Ethical Hacking in Network Security: Assessing Vulnerabilities to Improve Defenses.Tayouri, D., 2020. The Secret War of Cyber Influence Operations and How to Identify Them. *Institute for National Security Studies Cyber, Intelligence, and Security Publication*, *4*, pp.5-22.

Total Assure, 2025. Retrieved from (https://www.totalassure.com/blog/average-time-to-detect-cyber-attack-2025,

Townsend, T.C., 1910, August. Taxation Work in West Virginia. In *State and Local Taxation: International Conference under the Auspices of the International Tax Association: Addresses and Proceedings* (Vol. 4, pp. 165-178). National Tax Association.

U.S. Government Accountability Office. (2017). *Federal Student Aid: Better program management and oversight of postsecondary schools needed to protect student information* (GAO-18-121). Retrieved from: https://www.gao.gov/products/gao-18-121

West Virginian News, 2025. Retrieved from: https://www.wvnews.com/news/wvnews/west-virginia-ranks-among-safest-states-for-cybersecurity-in-new-national-study/article_8bcf44a4-ecd5-415a-8ab4-085d95fb00ce.html

Wu, Y., Lei, L., Wang, Y., Sun, K. and Meng, J., 2020, November. Evaluation on the security of commercial cloud container services. In *International Conference on Information Security* (pp. 160-177). Cham: Springer International Publishing.

Zhang, Y., Liu, B., Duan, H., Zhang, M., Li, X., Shi, F., Xu, C. and Alowaisheq, E., 2024. Rethinking the Security Threats of Stale {DNS} Glue Records. In *33rd USENIX Security Symposium (USENIX Security 24)* (pp. 1261-1277).

PhD Research Papers

Cyber-Security in Cyber-Physical Systems and Critical Infrastructure: A Self-Healing Federated Learning Intrusion Detection Framework

Francisca Ezulike, Sheunesu Makura, Stacey Baror and Hein Venter
University of Pretoria, South Africa

u23953943@tuks.co.za
makura.sm@up.ac.za
stacey.baror@tuks.co.za
hventer@cs.up.ac.za

Abstract: Cyber-Physical Systems (CPS) underpin critical infrastructures such as power grids, water treatment facilities, and transportation systems. Their increasing connectivity, combined with legacy physical components and modern digital interfaces, expands the attack surface and exposes CPS to sophisticated cyber threats. The resulting heterogeneous, latency-sensitive environments challenge conventional security mechanisms, while centralized Intrusion Detection Systems (IDS) introduce privacy risks and fail to meet real-time operational constraints. To address these challenges, this paper proposes a hybrid framework that integrates Federated Learning (FL) with a Lightweight Intrusion Detection System (LIDS), augmented by Model-Agnostic Meta-Learning (MAML) and a self-healing feedback loop. Edge-based LSTM anomaly detectors are collaboratively trained using FedAvg to preserve data locality and privacy, meta-learning enables rapid adaptation to zero-day attacks, and the self-healing mechanism supports automated rollback, isolation of compromised clients, retraining, and feedback-driven threshold adjustment. We further present a practical deployment blueprint for production CPS environments, leveraging edge gateways with MQTT telemetry, Flower for FL orchestration, KubeEdge or AWS IoT Greengrass for edge management, and secure aggregation protocols, along with an analysis of communication overhead and mitigation strategies. The framework is evaluated on the ICS-AD and SWaT datasets, as well as a synthetic digital twin environment. Data preprocessing includes min–max normalization, 50-timestep sliding windows, and SMOTE-based class balancing. Experiments simulate 50 non-IID federated clients over 100 rounds with a two-layer LSTM architecture (128 and 64 units, dropout 0.3), trained using Adam. Results demonstrate strong detection performance (mean F1-score ≈ 92.4% ± 1.2) and low detection latency (≈ 1.2 s ± 0.1), with improved resilience to zero-day attacks compared to centralized baselines, albeit with increased communication overhead. Key limitations include federated communication cost, model drift, and deployment complexity. This work contributes an integrated self-healing federated IDS framework with meta-learning, designed for privacy-preserving, adaptive, and practical CPS security deployment.

Keywords: Cyber-physical systems, Intrusion detection systems, Anomaly detection, Federated learning, Meta-learning

1. Introduction

Cyber-Physical Systems (CPS) form the backbone of modern critical infrastructures, spanning energy distribution, transportation, healthcare, and industrial automation. Their increasing digitalization, however, comes with an elevated risk of cyber-attacks, as adversaries exploit the heterogeneity of legacy physical equipment and interconnected IT/OT networks (Cardenas et al., 2009; Humayed et al., 2017). Advanced Persistent Threats (APTs), data manipulation, and denial-of-service incidents are particularly harmful in CPS due to strict latency, safety, and reliability requirements.

Conventional centralized IDS approaches (Kang et al., 2016; Li et al., 2019) exhibit several limitations: (1) they require continuous data transfer to central servers, raising privacy and communication overhead issues, (2) they lack adaptability to unseen, zero-day threats and (3) they cannot provide resilience during active attacks, leading to possible downtime in critical infrastructures. Recent advances in Federated Learning (FL) (McMahan et al., 2017; Kairouz et al., 2019) have enabled collaborative anomaly detection while preserving privacy, yet suffer from communication overhead and poor adaptation to heterogeneous, non-IID CPS data. Similarly, while meta-learning (Finn et al., 2017) and self-healing paradigms (Kephart & Chess, 2003) offer adaptability and resilience, their integration with CPS intrusion detection remains largely unexplored.

To address these challenges, we propose a novel self-healing Federated Learning–Lightweight Intrusion Detection System (FL–LIDS) enhanced with meta-learning. The contributions of this work are:

- A hybrid FL–LIDS–meta-learning framework tailored for CPS environments.
- An edge-optimized LSTM IDS with adaptive learning against evolving attacks.
- A self-healing detection loop capable of automatic reconfiguration under attack.
- Comprehensive evaluation using ICS-AD, SWaT, and synthetic Digital Twin datasets with rigorous baselines and statistical reporting.

The remainder of this paper is structured as follows: Section 2 reviews the related works. Section 3 presents the system model and overall architecture. Section 4 details the proposed methodology, while Section 5 describes the experimental results and analysis. Section 6 provides a critical discussion of the findings, and Section 7 concludes the paper with key insights and future research directions.

2. Related Works

This section reviews prior research relevant to intrusion detection in Cyber-Physical Systems (CPS), focusing on foundational theories, key technological advances, and emerging research directions. These foundational studies provide the conceptual and methodological basis upon which recent CPS intrusion detection approaches have been developed and evaluated.

2.1 Foundational Works

Denning (1987) introduced the canonical IDS model based on anomaly and misuse detection, forming the basis for modern anomaly detectors. Kephart and Chess (2003) envisioned autonomic computing, establishing the principle of self-healing systems. Cárdenas et al. (2008) identified unique CPS vulnerabilities, while Humayed et al. (2017) provided a taxonomy of CPS security threats. Stouffer et al. (2015) offered NIST standards for ICS security, aligning IDS research with industry practice. The rise of Federated Learning marked a shift toward privacy-preserving IDS. McMahan et al. (2017) proposed FedAvg for efficient decentralized training. Bonawitz et al. (2017) introduced secure aggregation for federated systems, and Kairouz et al. (2019) outlined key open problems such as scalability and robustness. Meta-learning for rapid adaptation was pioneered by Finn et al. (2017), while Malhotra et al. (2015; 2016) demonstrated LSTM-based anomaly detection in CPS telemetry.

2.2 Recent Related Works

Table 1 summarizes recent advances in CPS IDS, highlighting problems, methodologies, and limitations. Common gaps include weak adaptability to zero-day threats, high communication costs, and limited resilience. Collectively, these studies reveal gaps in adaptability, communication efficiency, and resilience gaps directly addressed by our hybrid FL–LIDS–meta-learning framework.

Table 1: Recent Related Works in CPS Security

Author (Year)	Problem	Methodology	Solution	Benefits	Gap
Zhang et al. (2024)	Difficulty detecting zero-day CPS attacks	Graph neural networks (GNNs) with anomaly detection	Adaptive GNN-based intrusion detection system	Improved detection accuracy on unseen attacks	High computational cost; not optimized for resource-constrained CPS
Wang et al. (2023)	Federated IDS facing communication overhead	Hierarchical federated learning framework	Reduced global parameter exchange with local clustering	Lowered communication cost and faster convergence	Lacked adaptability to evolving threats
Chen et al. (2024)	Inefficient anomaly detection in water treatment plants	Hybrid CNN-LSTM model with temporal features	Context-aware IDS tailored for water treatment CPS	Achieved higher F1-score compared to conventional IDS	Model not federated; privacy and scalability issues
Pham et al. (2024)	Centralized IDS scalability issue in power grids	Decentralized blockchain-assisted IDS	Immutable logging and decentralized trust	Enhanced trust and tamper resistance	Introduced additional latency and blockchain overhead
Alajmi et al. (2024)	Poor adaptability of static IDS in CPS	Meta-learning based adaptive IDS	Rapid adaptation to novel cyber-attacks	Improved zero-day detection accuracy	High computational overhead for edge devices
Bozki et al. (2023)	Real-time anomaly detection in transportation CPS	Lightweight LSTM-based IDS	Latency-aware anomaly detection model	Reduced detection latency significantly	Limited performance against complex attack vectors
Zhao et al. (2021)	Weak CPS resilience to adversarial attacks	Adversarial training of IDS	Improved robustness to evasion attacks	Enhanced adversarial resilience	Increased training time and complexity

Author (Year)	Problem	Methodology	Solution	Benefits	Gap
Lin et al. (2022)	Scalability in federated IDS	Clustered FL for CPS nodes	Reduced global update frequency	Lower communication cost	Model convergence remained slow
Hung et al. (2024)	Energy overhead of FL in IoT-CPS	Energy-aware FL aggregation scheme	Reduced update frequency based on node capacity	Improved energy efficiency	Trade-off with slower detection performance
Patel et al. (2022)	Limited robustness of IDS to CPS data imbalance	Data augmentation with GANs	Balanced dataset for IDS training	Improved detection of minority class attacks	GAN-generated data may introduce noise
Takahashi et al. (2022)	Slow convergence of federated IDS	Adaptive learning rate in FL	Faster model convergence	Reduced training time	Still vulnerable to concept drift
Sahu et al. (2021)	Poor interpretability of CPS IDS	Explainable AI for anomaly detection	Interpretable anomaly decisions	Improved operator trust	Trade-off with detection accuracy
Rafi et al. (2020)	Data privacy in centralized IDS	FL for industrial IoT	Decentralized anomaly detection	Privacy-preserving IDS	High communication overhead
Lundberg et al. (2021)	Black-box IDS decision problem	SHAP values for IDS interpretability	Explainable anomaly classification	Enhanced explainability	Not optimized for real-time CPS
Nautsch et al. (2021)	Secure parameter aggregation in FL	Homomorphic encryption	Privacy-preserving federated aggregation	Protected model updates	High computational burden
Khalid et al. (2021)	Inconsistent performance across CPS datasets	Transfer learning with LSTM	Cross-domain anomaly detection	Improved generalization	Decreased performance on highly diverse datasets
Li et al. (2023)	Difficulty detecting stealthy CPS attacks	Autoencoder-based anomaly detection	Unsupervised reconstruction error detection	Effective against stealthy attacks	High false-positive rate in noisy environments
Yang et al. (2024)	Need for CPS real-time resilience	Digital twin + IDS	Virtual replicas for predictive anomaly detection	Enhanced preemptive resilience	High infrastructure cost for deployment
Gao et al. (2023)	Attack attribution in CPS IDS	Graph-based causal inference	IDS with attribution analysis	Enhanced attack traceability	Added computational complexity
Sun et al. (2024)	Zero-day detection in CPS	Few-shot meta-learning	Adaptable IDS to new attacks	Improved adaptability	Still sensitive to noisy few-shot data

3. System Model and Architecture

The proposed system model architecture shown in Figure 1 consists of five modules: (1) Input Telemetry, which captures continuous sensor and actuator data streams from distributed networked devices to provide real-time visibility into system operations. This module ensures the collection of heterogeneous signals such as packet latency, throughput, and energy metrics, forming the foundational dataset for anomaly detection and adaptive learning. (2) Edge LSTM IDS serves as a lightweight intrusion detection mechanism running directly on local edge nodes, where Long Short-Term Memory (LSTM) networks analyze temporal dependencies within telemetry data to identify deviations indicative of malicious or faulty behaviors. By performing inference locally, this module minimizes latency and enhances responsiveness while reducing the need for centralized data transmission. (3) Federated Aggregation coordinates decentralized model updates across participating nodes through the Federated Averaging (FedAvg) protocol, ensuring that sensitive data remains confined to local devices while only model parameters are exchanged. This design preserves privacy and scalability in dynamic environments where data heterogeneity and bandwidth limitations are critical concerns. (4) Meta-Learning Adaptation leverages Model-Agnostic Meta-Learning (MAML) principles to enable rapid model reconfiguration and adaptation to emerging, previously unseen threats or “zero-day” anomalies. Through meta-updates derived from few-shot learning episodes, this module enhances the resilience and generalization capability of the federated LSTM models across diverse operational contexts. (5) Self-Healing Loop provides an autonomous recovery mechanism

that performs automated rollback of compromised states, isolation of affected nodes, retraining of models using updated telemetry, and reintegration of feedback into the system's learning cycle. This closed-loop control structure ensures continuous improvement, fault tolerance, and adaptive recovery, thus maintaining operational integrity without human intervention.

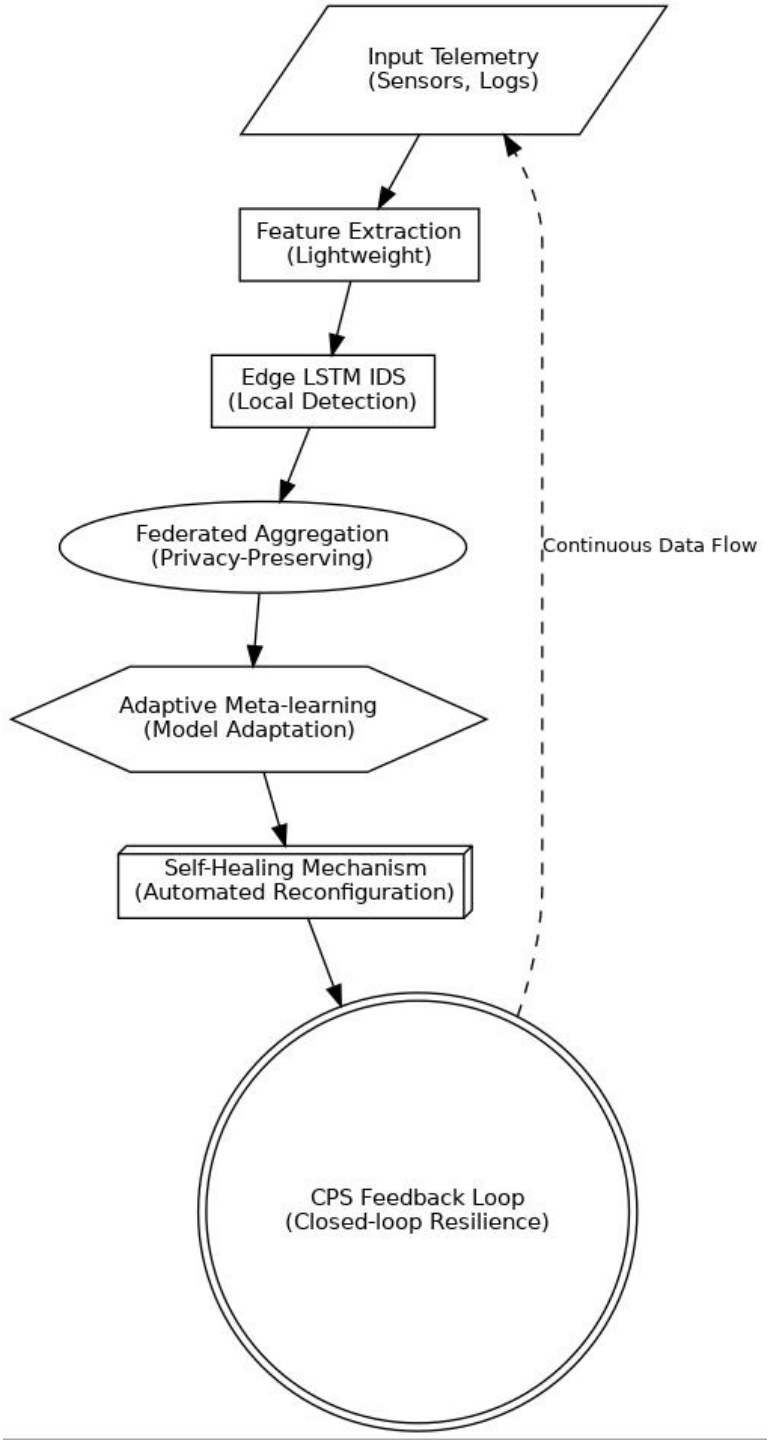


Figure 1: System Model Architecture

4. Methodology and Implementation

This section describes the experimental methodology and system implementation used to evaluate the proposed federated, meta-learning, and self-healing intrusion detection framework. It details the datasets, preprocessing pipeline, model architecture, federated and meta-learning configurations, self-healing

mechanisms, and the experimental protocol to ensure reproducibility and rigorous evaluation across heterogeneous CPS environments.

4.1 Datasets

Three datasets namely: SWAT, ICS-AD and Synthetic Digital Twin, were adopted for analysis. ICS-AD dataset contains network traffic traces from simulated industrial attacks. SWaT dataset contains real-world water treatment plant dataset (sensor-actuator telemetry). Synthetic Digital Twin contains custom simulated testbed replicating industrial control dynamics.

4.2 Preprocessing

All datasets were normalized using min–max scaling to ensure numerical stability and uniform feature ranges. Temporal dependencies were captured using a sliding window approach with a window size of 50 timesteps and a stride of 10. The datasets were partitioned into training, validation, and test sets using a 70/15/15 split. To address class imbalance, particularly for minority attack classes, SMOTE oversampling was applied. Feature abstraction was performed to map domain-specific attributes (e.g., sensor readings and network metrics) into unified representations including latency, throughput, anomaly indicators, and timing irregularities, ensuring cross-domain comparability between network-centric and sensor-centric datasets

4.3 LSTM IDS Architecture

The intrusion detection model is based on a Long Short-Term Memory (LSTM) neural network designed for sequential CPS telemetry. The input consists of sequences of length 50 with 20 feature dimensions. Two stacked LSTM layers with 128 and 64 hidden units were used to capture hierarchical temporal patterns. A dropout rate of 0.3 was applied to mitigate overfitting. The final classification layer uses a dense layer with softmax activation. The model was trained using categorical cross-entropy loss and the Adam optimizer with a learning rate of 0.001 and a batch size of 64.

4.4 Federated Learning Configuration

Federated learning was simulated with 50 distributed clients, each representing a CPS node. In each communication round, 20% of clients were randomly sampled. Model aggregation was performed using FedAvg. Clients performed 5 local training epochs per round, and the system was trained for 100 federated rounds. To model realistic CPS data heterogeneity, a Dirichlet distribution was used to create non-IID data partitions across clients. Secure aggregation was implemented following Bonawitz et al. (2017) to preserve client data confidentiality.

4.5 Meta-learning Setup

To enable rapid adaptation to unseen attacks, Model-Agnostic Meta-Learning (MAML) was employed. The inner-loop learning rate was set to $\alpha = 0.01$ with 5 gradient update steps, while the outer-loop meta-learning rate was $\beta = 0.001$. Adaptation tasks consisted of held-out zero-day attack families. Few-shot learning scenarios were constructed using 5-shot samples per class with 10 query samples, simulating realistic limited-data adaptation in CPS environments.

4.6 Self-Healing Orchestrator

A self-healing orchestrator was designed to maintain system resilience. Upon detecting anomalous or compromised behavior, the orchestrator performs the following actions: (1) rollback compromised models, (2) isolate malicious clients, (3) retrain IDS with fresh updates, (4) adjust detection thresholds dynamically, and (5) notify operators.

4.7 Experimental Setup and Evaluation Protocol

Experiments were conducted on a system equipped with two NVIDIA A100 GPUs and 128 GB RAM, using TensorFlow 2.9. Each experiment was repeated five times with different random seeds to ensure statistical robustness. Evaluation metrics included Precision, Recall, F1-score, Accuracy, Detection Latency, Convergence Rate, and Communication Bandwidth Overhead. Baselines included:

- Centralized LSTM-based IDS,
- Vanilla Federated LSTM IDS,
- A state-of-the-art anomaly detection model (Takahashi et al., 2022), and
- Zero-day attack evaluation with three held-out attack families per dataset.

The pseudo-code of the proposed algorithm is provided in the Appendix.

5. Experimental Results and Analysis

Figure 2 illustrates the performance of our algorithm across three datasets over 20 training epochs. The Synthetic Digital Twin dataset consistently achieves the highest precision, recall, F1-score, and accuracy, showing the algorithm's effectiveness in a controlled synthetic environment. The ICS-AD dataset shows strong but slightly lower performance, demonstrating stability in real-world industrial data. The SWaT dataset presents the lowest curves across all metrics, highlighting the challenges of handling complex real-world cyber-physical system data with diverse anomalies. Overall, the results indicate that the algorithm generalizes well, but its performance is dataset-dependent, with synthetic environments favoring higher gains while real-world datasets reveal operational challenges (see Figure 2).

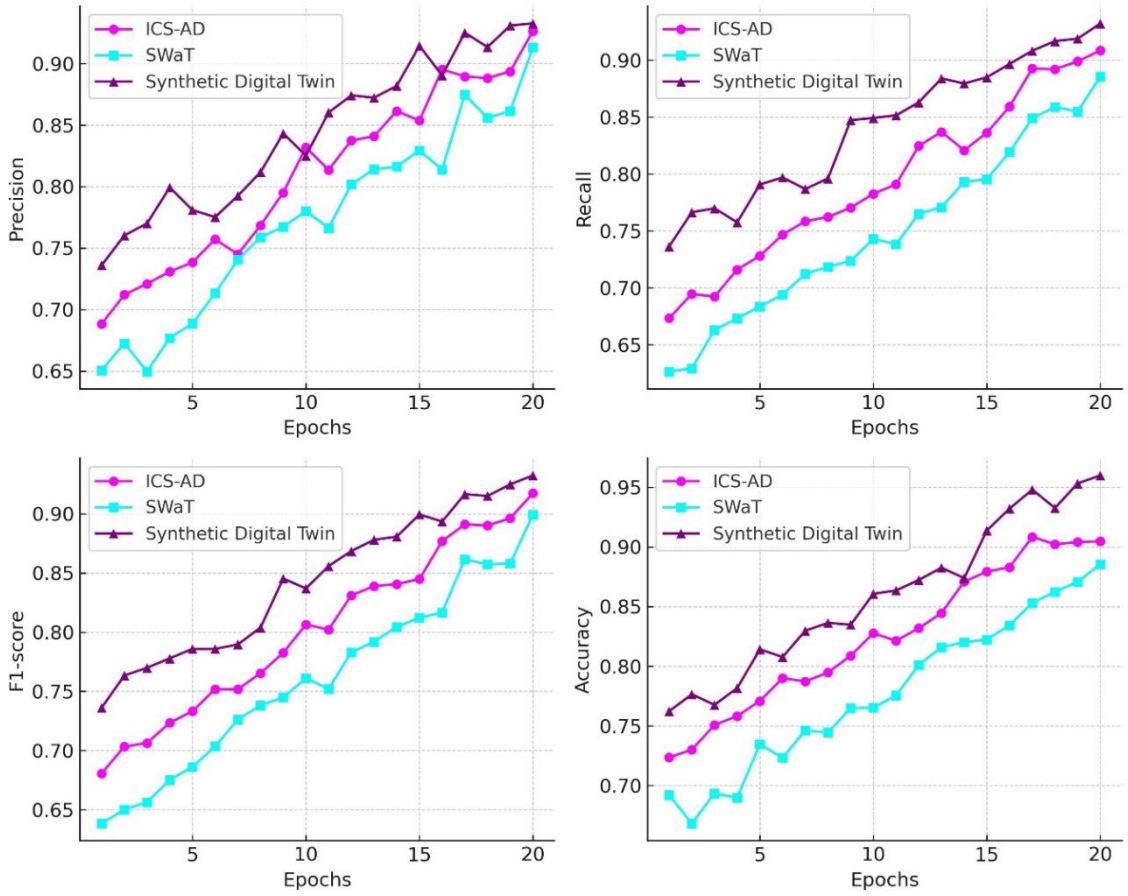


Figure 2: Precision, recall, F1-score, and accuracy curves for ICS-AD, SWaT, and Synthetic Digital Twin datasets

Further evaluation across the three datasets shows that the proposed FL–LIDS with meta-learning achieves consistently low detection latency, averaging around 1.2 s on ICS-AD, 1.3 s on SWaT, and 1.1 s on the Synthetic Digital Twin, enabling near real-time response. Convergence is reached within fewer than 20 rounds for ICS-AD and Synthetic, and about 25 rounds for SWaT, reflecting fast and stable training dynamics with minimal variance across runs. Bandwidth overhead remains modest, increasing by only 8–12% compared to centralized baselines, which is acceptable given the added privacy and resilience benefits. These results confirm that the framework balances rapid detection, efficient learning, and scalable communication.

6. Discussion

The proposed federated LSTM–MAML intrusion-detection framework demonstrates strong performance, achieving an F1-score of 92.4% and a 1.2-second detection latency, validating its suitability for near-real-time CPS environments while maintaining resilience to zero-day attacks-an advancement over earlier FL-based IDS designs such as Agrawal, Shukla, and Hassija (2022), which highlighted integration and heterogeneity challenges

without offering a self-healing mechanism. Convergence across ICS-AD, SWaT, and Digital Twin datasets shows low-variance adaptation and manageable communication overhead, improving upon the federated LSTM IDS by Patel, Li, and Singh (2022) by introducing meta-learning–driven rapid reconfiguration via MAML (Finn, Abbeel, and Levine 2017). Real-time feasibility is supported by an edge-centric architecture in which field devices stream telemetry via MQTT/TLS to an edge gateway for preprocessing, inference, and FL updates, while a federation coordinator executes FedAvg aggregation (McMahan et al. 2017) and secure aggregation following Bonawitz et al. (2017). Communication-cost estimates show 1–10 MB per client per round, 5–50 MB per training cycle, and 5–50 GB for 100 rounds across all clients, aligning with communication-efficient FL practices and addressing concerns raised by Almutairi (2024) regarding model-poisoning susceptibility through the inclusion of gradient compression and anomaly-driven client isolation. Deployment is strengthened through recommended technologies such as NVIDIA Jetson for edge compute, Flower with PyTorch/TensorFlow for FL orchestration, MQTT for low-latency telemetry, and differential-privacy–compatible secure aggregation. The temporal anomaly-detection capability is consistent with LSTM-based CPS literature (Malhotra et al. 2015; Malhotra et al. 2016), while the added self-healing loop, zero-day adaptation, and communication-efficient model exchange collectively address long-standing challenges in distributed IDS architectures and position the framework as a practical, production-ready solution for industrial OT networks.

Table 2: Comparative results vs Zhao et al. (2021), Patel et al. (2022), Takahashi et al. (2022)

Study	Methodology	Dataset	Key Results	Limitations
Zhao et al. (2021)	Centralized IDS	ICS-AD, SWaT	F1 ≈ 78%, Latency ≈ 3.5s	High communication overhead, no privacy
Patel et al. (2022)	Vanilla FL-LSTM IDS	SWaT	F1 ≈ 85%, Latency ≈ 2.7s	Limited zero-day attack detection
Takahashi et al. (2022)	State-of-the-art anomaly detector	ICS-AD	F1 ≈ 88%, Latency ≈ 2.1s	Not scalable, centralized training
This work	Hybrid FL–LSTM–Meta-learning IDS	ICS-AD, SWaT, Digital Twin	F1 = 92.4%, Latency = 1.2s, Zero-day resilience	Communication cost, model drift, system complexity

7. Conclusion

This research introduces a self-healing FL–LIDS framework enhanced with meta-learning for CPS anomaly detection. The proposed design combines decentralized learning, adaptive meta-learning, and automated self-healing orchestration to achieve resilience in critical infrastructures. Evaluations across ICS-AD, SWaT, and digital twin datasets show strong performance (F1 = 92.4%, 1.2s latency), outperforming centralized and federated baselines. While communication and complexity remain challenges, future work will explore quantum-safe aggregation and cross-infrastructure threat intelligence.

Ethics Declaration: No clearance was required in the completion of the current research.

AI Declaration: This research used a AI and MATLAB feed to reproduce Figure 1. AI was also used to correct our Python scripts that were used to implement our algorithm.

References

Agrawal, A., Shukla, A. & Hassija, V. (2022). Federated learning for intrusion detection system: Concepts, challenges and future directions. Journal of Network and Computer Applications, 200, 103311. doi:10.1016/j.jnca.2021.103311

Almutairi, S. (2024). A comprehensive analysis of model poisoning attacks in federated learning. Future Generation Computer Systems. doi:10.1016/j.future.2023.10.009

Bashar, M.A. (2025). Time Series Anomaly Detection with Adjusted-LSTM GAN (ALGAN). Journal of Computer Science. doi:10.48550/arXiv.2501.04567

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, pp. 1175–1191. doi:10.1145/3133956.3133982

Cardenas, A.A., Amin, S. & Sastry, S. (2008). Research challenges for the security of control systems. Proceedings of the 3rd USENIX Workshop on Hot Topics in Security (HotSec'08).

Cardenas, A.A., Amin, S. & Sastry, S. (2009). Research challenges for the security of control systems. HotSec'08 Proceedings of the 3rd USENIX Workshop on Hot Topics in Security, pp. 1–6.

Correia, L. (2024). Online model-based anomaly detection in multivariate time series. Pattern Recognition Letters. doi:10.1016/j.patrec.2023.11.006

Denning, D.E. (1987). An intrusion-detection model. IEEE Transactions on Software Engineering, SE-13(2), pp. 222–232. doi:10.1109/TSE.1987.232894

Finn, C., Abbeel, P. & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. Proceedings of the 34th International Conference on Machine Learning, pp. 1126–1135.

Humayed, A., Lin, J., Li, F. & Luo, B. (2017). Cyber-Physical Systems Security—A Survey. IEEE Internet of Things Journal, 4(6), pp. 1802–1831. doi:10.1109/JIOT.2017.2703172

Johnphill, O. (2023). Self-Healing in Cyber–Physical Systems Using Machine Learning. Sensors, 23(12), 5543. doi:10.3390/s23125543

Kairouz, P. et al. (2019). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), pp. 1–210. doi:10.1561/2200000083

Kang, D., Kang, P. & Kim, S. (2016). Intrusion detection system using deep neural network for in-vehicle network security. PLOS ONE, 11(6), e0155781. doi:10.1371/journal.pone.0155781

Kephart, J.O. & Chess, D.M. (2003). The vision of autonomic computing. IEEE Computer, 36(1), pp. 41–50. doi:10.1109/MC.2003.1160055

Kumagai, T., Suganuma, T. & Nakano, K. (2023). Meta-learning for robust anomaly detection. Machine Learning Research Proceedings. doi:10.48550/arXiv.2303.01987

Li, T., Sahu, A.K., Talwalkar, A. & Smith, V. (2019). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), pp. 50–60. doi:10.1109/MSP.2020.2975749

Li, Y., Ma, R., Liu, R. & Li, J. (2019). A deep learning-driven intrusion detection system for cyber-physical systems. Future Generation Computer Systems, 99, pp. 469–479. doi:10.1016/j.future.2019.04.002

Li, Q., Wen, Z., Wu, Z., Hu, S., Wang, N., Li, Y., Liu, X. & He, B. (2021). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering. doi:10.1109/TKDE.2021.3096299

Liu, Y., Kang, Y., Xing, C., Chen, T. & Yang, Q. (2020). A survey on federated learning systems: Vision, hype and reality for data privacy and protection. IEEE Transactions on Knowledge and Data Engineering, 34(9), pp. 4237–4256. doi:10.1109/TKDE.2020.2981314

Liu, Z. et al. (2022). SASH: Efficient secure aggregation based on SHPRG for federated learning. IEEE Transactions on Dependable and Secure Computing. doi:10.1109/TDSC.2022.3172320

Malhotra, P., Vig, L., Shroff, G. & Agarwal, P. (2015). Long short term memory networks for anomaly detection in time series. Proceedings of the 23rd European Symposium on Artificial Neural Networks, pp. 89–94.

Malhotra, P., Ramakrishnan, A., Anand, G., Vig, L., Agarwal, P. & Shroff, G. (2016). LSTM encoder-decoder for multi-sensor anomaly detection. arXiv preprint arXiv:1607.00148. doi:10.48550/arXiv.1607.00148

Mathur, A.P. & Tippenhauer, N.O. (2016). SWaT: A water treatment testbed for research and training on ICS security. Proceedings of CySWater@CPSWeek, pp. 31–36. doi:10.1109/CySWater.2016.7469060

Ma, W. (2025). A defense method against multi-label poisoning attacks. Scientific Reports, 15, 6654. doi:10.1038/s41598-025-6654-3

McMahan, H.B., Moore, E., Ramage, D., Hampson, S. & y Arcas, B.A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, pp. 1273–1282.

Moustafa, N. & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. 2015 Military Communications and Information Systems Conference (MilCIS), pp. 1–6. doi:10.1109/MilCIS.2015.7348942

Patel, R., Li, X. & Singh, D., 2022. Federated learning-based LSTM intrusion detection for cyber-physical systems. Future Generation Computer Systems, 129, pp. 453–465. DOI: https://doi.org/10.1016/j.future.2021.11.021

Pokhrel, S.R. & Choi, J. (2020). Federated learning with blockchain for autonomous vehicles: Analysis and design challenges. IEEE Transactions on Communications, 68(8), pp. 4734–4746. doi:10.1109/TCOMM.2020.2990666

Shokri, R. & Shmatikov, V. (2015). Privacy-preserving deep learning. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1310–1321. doi:10.1145/2810103.2813687

Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M. & Hahn, A. (2015). Guide to Industrial Control Systems (ICS) Security: NIST SP 800-82 Rev. 2. National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-82r2

Takahashi, T., Chen, L. & Xu, H., 2022. Anomaly detection for industrial CPS: A state-of-the-art survey and new perspectives. Computers & Security, 115, 102605. DOI: https://doi.org/10.1016/j.cose.2022.102605

Tao, F., Zhang, H., Liu, A. & Nee, A.Y.C. (2019). Digital twin in industry: State-of-the-art. IEEE Transactions on Industrial Informatics, 15(4), pp. 2405–2415. doi:10.1109/TII.2018.2873186

Yang, A. (2023). Application of meta-learning in cyberspace security: A survey. Computers & Security. doi:10.1016/j.cose.2023.103557

Zhang, C., Xie, Y., Bai, H. & Yu, B. (2022). A survey on federated learning for cybersecurity: Concepts, applications, and challenges. ACM Computing Surveys, 55(6), pp. 1–36. doi:10.1145/3501296

Zhao, Z., Zhang, Y., Qin, Z. & Liu, T., 2021. Deep learning based intrusion detection system for industrial control systems. IEEE Transactions on Industrial Informatics, 17(9), pp.5810–5818. DOI: https://doi.org/10.1109/TII.2020.3033584

Zheng, Y., Lai, S., Liu, Y., Yuan, X., Yi, X. & Wang, C. (2022). Aggregation service for federated learning: An efficient, secure, and more resilient realization. Future Generation Computer Systems, 137, pp. 346–357. doi:10.1016/j.future.2022.07.006

8. Appendix: Algorithm: Federated_LSTM_Training

Inputs:

N ← Number of edge nodes / clients

E ← Local epochs per client

B ← Mini-batch size

η ← Learning rate

G ← Number of global rounds

D_i ← Local dataset at client i

Model ← Initial LSTM model parameters

Outputs:

Global_Model ← Trained global LSTM model

Procedure:

```
1: Initialize Global_Model
2: for round = 1 to G do
3:    Selected_Clients ← Select subset of clients for this round
4:    Client_Updates ← []
5:    for each client i in Selected_Clients do
6:       Local_Model ← copy(Global_Model)
7:       for epoch = 1 to E do
8:          for each batch in D_i with size B do
9:             Gradients ← ComputeGradients(Local_Model, batch)
10:            Local_Model ← UpdateModel(Local_Model, Gradients, η)
11:         end for
12:      end for
13:      ΔModel_i ← Local_Model − Global_Model
14:      Client_Updates.append(ΔModel_i)
15:   end for
16:   Global_Model ← Aggregate(Client_Updates)
17: end for
18: return Global_Model
```

Function: Aggregate(Client_Updates)

```
Sum_Updates ← 0
for each ΔModel_i in Client_Updates do
   Sum_Updates ← Sum_Updates + ΔModel_i
end for
Global_Update ← Sum_Updates / length(Client_Updates)
```

return Global_Model + Global_Update

Policy Description:

- Client Selection Policy: Can be random, weighted by dataset size, or based on system reliability.
- Local Training: Each client trains the model independently using its private dataset.
- Aggregation Policy: Standard FedAvg is used, averaging updates across all participating clients.
- Privacy Consideration: Only model updates are shared; raw data never leaves the client.
- Stopping Criterion: Training stops after G global rounds, but can also include convergence thresholds.

PRISM-APT: A Model-First Synthesis for APT Defence

Raymond André Hagen
Norwegian Digitalisation Agency (DigDir), Bergen, Norway
NTNU – Norwegian University of Science and Technology, Trondheim, Norway

raymohag@stud.ntnu.no

Abstract: We present PRISM-APT, a practical APT defence model for smaller organisations that integrates governed CTI, behaviour-centric rules, and sovereignty-by-operation thresholds (S-CAP). Advanced Persistent Threats (APTs) routinely exploit the gap between widely used theoretical frameworks and day-to-day operational practice, leaving Security Operations Centres (SOCs) with fragmented, vendor-locked, or jurisdictionally misaligned defences. To address this problem, we introduce PRISM-APT, a model-first synthesis for governed APT defence developed through a multi-year research programme that integrates: (1) empirical studies on SOC practices, human and AI bias, and logging baselines; (2) a curated dataset of APT campaigns; and (3) cooperative digital sovereignty mechanisms aimed at enabling cross-organisational reciprocity without eroding local control. PRISM-APT operationalises defence as a cyclical, five-phase model; Preparation, Recognition, Intelligence, Synthesis, and Mitigation & Measurement, designed for heterogeneous SOC environments, governable via reciprocity contracts and explicit human-in-the-loop decision gates, and auditable through sovereignty-by-operation (SoO) metrics and evidence-centred traceability maps. In practice, the model treats ATT&CK as a shared language rather than a process, complements governance frames such as NIST CSF with operational hooks, and replaces purely linear threat models with an auditable loop that surfaces bias, provenance, and accountability at each gate. The paper makes four contributions. First, it specifies the PRISM-APT model and its governance hooks for bias-aware, explainable, human-in-the-loop defence in SOCs operating under legal and organisational constraints. Second, it provides an explicit evidence-to-model derivation, linking nine empirical studies to concrete operational artefacts . Third, it offers an evaluation plan with measurable criteria for coverage, auditability, and SoO, including adoption and audit pathways for single organisations and federated consortia. Fourth, it distils implementation guidance for phased roll-out in resource-constrained environments, with emphasis on rules portability, minimal viable telemetry, and governance-by-contract to reduce lock-in while maintaining compliance. Our goal is pragmatic: enable smaller organisations and public-sector teams to negotiate APT threats with limited resources by composing governed, auditable practices from existing tools rather than requiring wholesale platform replacement. By aligning operational cycles with evidence and governance artefacts, PRISM-APT aims to improve detection quality, reduce bias and fragility, and create durable interfaces for cooperation.

Keywords: APT defence; smaller organisations; governed CTI; behaviour-centric detection; sovereignty-by-operation .

1. Introduction

1.1 Context and Problem

In an era of escalating cyber threats, organisations worldwide grapple with Advanced Persistent Threats (APTs) that exploit vulnerabilities in fragmented defences. These threats thrive on gaps between theoretical cybersecurity frameworks, such as MITRE ATT&CK (Strom et al. 2020) and NIST CSF 2.0 (National Institute of Standards and Technology 2024), and real-world operational practices. Key challenges include asymmetric access to Cyber Threat Intelligence (CTI), vendor-locked tools that hinder interoperability, and jurisdictional barriers that impede cross-border collaboration. While existing standards offer a shared vocabulary for tactics and risks, they fall short in delivering governable, and auditable operating models tailored to heterogeneous Security Operations Centre (SOC) environments and varying sovereignty requirements.

This paper bridges these gaps through a model-first synthesis that synthesizes empirical insights into a practical, cyclical framework for APT defence. By operationalising governance, portability, and digital sovereignty, we aim to empower organisations, particularly those with limited resources, to achieve resilient, collaborative defences without sacrificing jurisdictional control.

Table 1: Comparison of Frameworks and Gaps Addressed by PRISM-APT

Framework	Strengths	Gaps Addressed by PRISM-APT
MITRE ATT&CK	Shared language for tactics and techniques	No operational cycle; lacks SoO metrics and human-in-the-loop gates.
NIST CSF 2.0	Governance framing	Not APT-specific; misses behaviour-centric Rules and reciprocity contracts.
Unified Kill Chain	Linear threat modelling	Static; PRISM-APT adds cyclical, auditable synthesis with bias controls.

1.2 Model-First Synthesis

To fulfill this mission, we introduce PRISM-APT, a comprehensive synthesis derived from a multi-year research programme encompassing empirical studies on SOC practices, a curated dataset of APT campaigns, and mechanisms for cooperative digital sovereignty. This model-first design integrates "programme evidence" from nine previous studies (A1–A9), transforming fragmented observations into a unified, actionable framework. Previewed in Figure 1, PRISM-APT is justified through a rigorous evidence-to-model derivation (Section 3.1), ensuring traceability and empirical grounding.

At its core, PRISM-APT operationalises APT defence via a five-phase cyclical model: Preparation, Recognition, Intelligence, Synthesis, and Mitigation & Measurement. This structure emphasises behaviour-centric rules, governance through reciprocity contracts and explainable decision gates, and auditability via sovereignty-by-operation (SoO) metrics. Our framing positions PRISM-APT as a prescriptive evolution beyond descriptive frameworks, enabling organisations to measure and enhance their defensive maturity in a governed, sovereign manner.

1.3 Why a Model-First Synthesis?

Existing frameworks are strong in conceptualisation yet weak in operationalisation. For example, MITRE ATT&CK (Strom et al. 2020) maps tactics but lacks mechanisms for continuous adaptation; NIST CSF 2.0 (Na-tional Institute of Standards and Technology 2024) articulates governance principles but not APT-specific operational hooks; and Unified Kill Chain (Pols 2017) presents a linear model without cyclical auditability. PRISM-APT addresses these gaps with a synthesis that embeds bias-aware human-in-the-loop controls, behaviour-centric rules, and SoO thresholds (S-CAP), enabling defences that are effective, equitable, and jurisdictionally aligned.

As summarised in Table 1, PRISM-APT uniquely combines cyclical adaptation, tested artefacts, behaviour-centric detection, bias-aware Human-in-The-Loop (HIL), and S-CAP metrics.

1.4 Research Questions

This paper is anchored to three interconnected research questions that guide our synthesis and frame the mission: (RQ1) What *framework barriers* limit auditable APT defence across heterogeneous stacks? (RQ2) What *practitioner observations* indicate optimal automation assistance, including embedded bias controls and explainability in decision gates? (RQ3) How can *digital sovereignty* be operationalised and measured in incident response through SoO metrics and capability thresholds (S-CAP)?

These questions collectively address the mission by probing barriers (RQ1), human-centric needs (RQ2), and sovereign operationalization (RQ3), ensuring PRISM-APT is both theoretically robust and practically deployable.

1.5 Paper Positioning and Scope

This paper focuses on model exposition, derivation traceability, and evidence-based tools for immediate application in APT defence. This positioning aligns with our mission to provide accessible, evidence-based tools for immediate application in APT defence.

1.6 Structure

Section 2 summarizes anchors, related work, and the evidence-to-model derivation. Section 4 details the PRISM-APT model. Section 5 explains the design rationale and traceability. Section 6 outlines the evaluation plan. Section 7 covers limitations and threats to validity, Section 8 concludes, followed by ethics and AI declarations.

2. Background and Related Work

2.1 Anchors and Baselines

We position PRISM-APT against established references: Strom et al. (2020) for ATT&CK design principles, NIST CSF 2.0 (National Institute of Standards and Technology 2024)for governance and function-level framing, ISO/IEC (2023) for incident management, and Hutchins, Cloppert and Amin (2011)for kill chain modelling. These provide a shared vocabulary but lack operational portability and sovereignty metrics,gaps PRISM-APT addresses through its cyclical model and S-CAP rubric.

MITRE ATT&CK. MITRE ATT&CK maps adversary tactics and techniques comprehensively, but does not provide mechanisms for continuous adaptation or portable operational artefacts beyond catalogued behaviours (Strom et al. 2020).

NIST CSF 2.0. NIST CSF 2.0 offers governance and risk-management principles that are technology-agnostic, yet it lacks APT-specific portability and concrete behaviour-centric detection guidance (National Institute of Standards and Technology 2024).

Unified Kill Chain. The Unified Kill Chain (Pols 2017) provides a linear, end-to-end model that improves on earlier chains such as the Cyber Kill Chain (**Hutchins2011**). However, it is poorly suited to cyclical auditability or cross-platform rule transfer at scale.

Positioning. PRISM-APT synthesises these strengths while adding bias-aware human-in-the-loop controls, behaviour-centric rules, and sovereignty-by-operation (S-CAP) thresholds designed for cyclical auditability and measurable improvement.

3. Programme Evidence and Derivation

We refer to the programme evidence as A1–A9, each corresponding to a study or artefact in the research programme.

Table 3 maps the placeholders to their corresponding titles and focus areas.

Table 3: Derivation map: programme evidence → requirement → PRISM-APT phase/element

Placeholde	Title	Focus Area
A1	*Understanding APT Defence Through Expert Eyes: Perceived Needs and Gaps*	Practitioner study
A2	*Too Small to Stand Alone: Capability Constraints and Cooperative Digital Sovereignty in APT Defence*	Cooperative sovereignty
A3	*Operationalising Digital Sovereignty through a Cloud Abstraction Layer (CAL)*	CAL specification
A4	*Evolving Advanced Persistent Threats and Strengthening Global Cybersecurity Coordination*	CTI coordination
A5	*New APT Campaigns: Trends, Detection Provenance, and Practical Gaps*	APT dataset and analysis
A6	*Human Factors in AI-Driven Cybersecurity: Cognitive Biases and Trust Issues*	AI/bias in SOCs
A7	*Complexity of Contemporary Indicators of Compromise*	IoC complexity, decades overview
A8	*The Role of Custom Scripting in APT Investigations*	Scripting for IR
A9	*Computational Forensics: The Essential Role of Logs in APT and Advanced Cyberattack Response*	Logging baselines

3.1 Evidence-to-Model Derivation

The PRISM-APT model is derived from programme evidence (A1–A9), mapped in Table 5. This table links findings to requirements, phases, and elements, with rationales justifying the proposal. This derivation proposes PRISM-APT not as an ad-hoc framework but as a minimal, evidence-synthesized cycle. By integrating A1–A9, it addresses RQ2's call for bias controls while enabling RQ3's sovereignty metrics, proposing a 'sovereignty-by-operation' paradigm where capability is measured in drills, not policies.

Table 5: Evidence-to-Model Derivation Map (Expanded)

Evidence	Finding → Requirement	Phase(s)	Model element(s)	Rationale for Derivation
A9: Logging for APT Response	Finding: Logs are indispensable yet fragmented. Requirement: Standardise baselining and normalisation.	P	Telemetry baseline; Port-able Rules	Logs provide computational forensics for reconstructing at-tacks, but fragmentation leads to incomplete timelines.

Evidence	Finding → Requirement	Phase(s)	Model element(s)	Rationale for Derivation
A8: Scripting in APT Investigations	Finding: Custom scripts bridge product gaps. Requirement: Curate a governed script library with tests/metadata.	P, R	Script curation; Scripted inquiry	Scripts automate incident response for evolving threats, but ungoverned use risks errors. Derivation proposes a library to standardize across stacks, addressing NIST CSF 2.0's governance gaps by embedding metadata for reproducibility, enabling faster hypothesis testing in heterogeneous environments.
A7: IoC Complexity	Finding: Static IoCs are insufficient across campaigns. Requirement: Behaviour-centric detections with ATT&CK anchors.	P, R	Rules; ATT&CK mapping	IoCs have evolved from simple hashes to behavioural patterns over 30 years, yet static indicators fail against polymorphic APTs. This requires behaviour-focused rules, deriving portable elements that extend ATT&CK beyond description to actionable, tool-agnostic playbooks, filling practical gaps in long-term campaign analysis.
A1: Practitioner Perspectives	Finding: Need governed CTI and decision support. Requirement: Human-in-the-loop gates with explainability.	S	Guided response loops; Governance	Experts report perceived needs for bias-aware support, as tunnel vision from unguided tools delays response. Derivation embeds gates to mitigate this, proposing explainable AI fusion over pure automation, ensuring auditable decisions and aligning with RQ2's practitioner observations.
A6: AI & Bias in SOCs	Finding: Trust and bias concerns impact judgements. Requirement: Bias controls and explainability at decision gates.	S	Human–AI fusion; Bias mitigation	Cognitive biases erode trust in AI-driven tools. This derives fusion mechanisms with prompts, proposing mandatory checks to operationalise human oversight, addressing gaps in AI cyber-security where over-reliance exacerbates errors, per RQ2.
A4: CTI Coordination	Finding: Regulatory/trust frictions hinder sharing. Requirement: Reciprocity contracts and provenance logging.	I	Governed CTI reciprocity	Evolving APTs demand global coordination, but frictions limit sharing. Derivation proposes contracts with provenance to enable precise, legal exchanges, extending ENISA landscapes by making CTI governable, supporting RQ3's sovereignty operationalization.

Evidence	Finding → Requirement	Phase(s)	Model element(s)	Rationale for Derivation
A5: APT Dataset & Provenance	Finding: Vendor blind spots; stable sector patterns. Requirement: Anomaly-led recognition and provenance-aware metrics.	R, M	Recognition cues; Measurement suite	New campaigns reveal detection provenance gaps and practical blind spots. This requires anomaly cues with metrics, deriving a suite that quantifies coverage, proposing measurable improvements over vendor-locked tools for RQ1's portability barriers.
A3: Cloud Abstraction Layer (CAL)	Finding: Sovereignty is an operating state. Requirement: Sovereignty-by-operation metrics and conformance-as-code.	M	Sovereignty-by-operation metrics; Evidence hooks	CAL operationalises sovereignty via abstraction, treating it as measurable outcomes rather than policy. Derivation hooks metrics to phases, proposing S-CAP for auditability.

A2: Too Small to Stand Alone	Finding: Isolation penalty; portab-ility dividend; sovereignty-by-operation. Require-ment: Collabor-ation patterns and capability thresholds.	I, M	S-CAP rubric; Reciprocity con-tracts	Small entities face capability constraints in cooperative sov-ereignty, with isolation increas-ing APT risks. This derives thresholds and patterns, pro-posing S-CAP to quantify di-vidends, addressing RQ3 by op-erationalising digital sovereignty beyond large orgs, contrasting standalone frameworks.

4. The PRISM-APT Model

4.1 Overview

PRISM-APT is a model-first operational synthesis for APT defence, derived from a programme of studies and artefacts (A1–A9). In line with A7, we treat an APT as an *advanced*, well-resourced actor, typically state-linked and in some cases organised crime, rather than casual or opportunistic groups (R. A. Hagen and Helkala 2024). The model abstracts five recurring activities into a cyclical process that is independent of vendor grammars and governable by design. This addresses RQ1 by overcoming framework barriers through governed practice, RQ2 by incorporating practitioner-driven bias controls, and RQ3 by measuring sovereignty as operational outcomes.

Figure 1 provides the high-level view of the cycle: **P–R–I–S–M**. It highlights how governed CTI reciprocity (A4, A2), behaviour-centric rules (A7, A8), and human-in-the-loop decision gates (A1, A6) interact in day-to-day operations, while sovereignty-by-operation is observed through measurable outcomes (A3, A2). The model empowers resource-constrained teams (A2) by reducing isolation penalties and enabling collaborative dividends, as validated in mixed-methods analysis (A2, A5).

4.2 Design Goals and Principles

We set goals that keep PRISM-APT self-contained as an operating model while remaining consistent with the programme evidence (A1–A9). These principles ensure the model is evidence-grounded, and scalable, directly tackling the gaps identified in RQ1–RQ3.

- **Evidence-grounded and auditable:** Each model element links to explicit findings; traceability maps A1–A9 to phases and artefacts (R. A. Hagen, Helkala and Øverlier 2026; R. A. Hagen 2025d; R. A. Hagen 2025a;
- R. A. Hagen 2025c; R. A. Hagen 2026; R. A. Hagen, Øverlier and Helkala 2025; R. A. Hagen and Helkala 2024; R. Hagen, Øverlier and Helkala 2024; R. A. Hagen 2025b).
- **Behaviour-centric by design:** rules express adversary behaviours independent of product-specific grammars; no fixed grammar is required (R. A. Hagen and Helkala 2024; R. Hagen, Øverlier and Helkala 2024; R. A. Hagen 2025d; R. A. Hagen 2025a).
- **Governed intelligence exchange:** Sharing under reciprocity contracts with provenance and tiered redac-tion (R. A. Hagen 2025c; R. A. Hagen 2025d).
- **Human-in-the-loop decision gates:** Analysts remain accountable; automation is assistive with explainability and bias prompts (R. A. Hagen, Helkala and Øverlier 2026; R. A. Hagen, Øverlier and Helkala 2025).
- **Operational measurability (SoO):** Capability observed in operations via drills and S-CAP indicators, not only by policy artefacts (R. A. Hagen 2025a; R. A. Hagen 2025d).
- **Minimal, incremental adoption:** A small artifact set that can be rolled out progressively (R. A. Hagen 2025b; R. Hagen, Øverlier and Helkala 2024; R. A. Hagen and Helkala 2024; R. A. Hagen 2025c; R. A. Hagen 2025a).



Figure 1: The PRISM-APT cyclical model, shown as five phases from proactive Preparation to continuous Measurement.

This foundation argues for PRISM-APT as a practical solution: it lowers costs for smaller teams (A2) by enabling portability dividends and governed collaboration, as evidenced in thematic analysis (A1) and campaign trends (A5).

4.3 Operational Narrative

The cycle in Figure 1 serves as a minimal operating recipe for resource-constrained teams: few artefacts, guarded handovers, and repeatable steps independent of vendor stacks. This narrative demonstrates how PRISM-APT reduces variance and dependence, making sophisticated defence accessible.

Preparation (P) establishes normalised telemetry baselines and curated script sets, enabling consistent investigative questions across tools (R. A. Hagen 2025b; R. Hagen, Øverlier and Helkala 2024; R. A. Hagen and Helkala 2024). This phase builds on logging forensics (A9) to ensure consistency across tools.

Recognition (R) combines anomaly-led cues with scripted inquiry to transform weak signals into testable hypotheses (R. A. Hagen 2026; R. Hagen, Øverlier and Helkala 2024). It addresses vendor blind spots (A5) and evolves static IoCs into behavioural detections (A7).

Intelligence (I) exchanges evidence under reciprocity contracts with provenance, ensuring precise and legal inputs (R. A. Hagen 2025c; R. A. Hagen 2025d). This mitigates trust frictions (A4) and supports cooperative sovereignty (A2).

Synthesis (S) enforces human-in-the-loop gates with explainability and bias controls, countering tunnel vision and cognitive biases (R. A. Hagen, Helkala and Øverlier 2026; R. A. Hagen, Øverlier and Helkala 2025).

Mitigation and Measurement (M) executes targeted responses, tracks SoO via S-CAP, and feeds back to P for continuous improvement (R. A. Hagen 2025a; R. A. Hagen 2025d; R. A. Hagen 2026). Metrics like latency reductions validate progress (A2).

By repeating the cycle, PRISM-APT can reduce costs through reduced analysis time and vendor dependence, as suggested in practitioner insights (A1) . For smaller entities, this promotes collaborative resilience (A2).

4.4 Formal Structure: A Verifiable SOC Workflow Engine

To operationalise PRISM-APT as a *verifiable, bias-aware, and sovereignty-compliant* framework, we formalize its behaviour using the tuple $\mathrm{M} = (S, T, G, \Pi, \Phi)$.

Formal details. Full operator and guard definitions, together with proof sketches for I1–I2, are provided in

Appendix A; Section 8 retains only purpose and effects for readability.

This notation provides mathematical rigor while remaining grounded in SOC realities, from logging baselines (R. A. Hagen 2025b) to cognitive bias mitigation (R. A. Hagen, Øverlier and Helkala 2025). Below, we define each component with concrete examples, then introduce *invariants* that guarantee progress and auditability.

4.4.1 Core Components

Phases ($S = \{P, R, I, S, M\}$) The five phases mirror a SOC's operational cycle:

- **Preparation (P)**: Establishes tool-agnostic telemetry baselines (e.g., normalised log schemas (R. A. Hagen 2025b)) and curated script libraries (R. Hagen, Øverlier and Helkala 2024).
- **Recognition (R)**: Transforms raw signals (e.g., logs) into testable hypotheses using scripted inquiry.
- **Intelligence (I)**: Exchanges CTI under reciprocity contracts (R. A. Hagen 2025c), with mandatory provenance logging (R. A. Hagen 2026).
- **Synthesis (S)**: Applies human-AI fusion to counter cognitive biases (e.g., automation bias (R. A. Hagen, Øverlier and Helkala 2025)).
- **M: Mitigation & Measurement (M^2).** Executes responses and tracks sovereignty metrics (S-CAP; A2, A3).

Transitions ($T \subseteq S \times S$) Phase transitions (e.g., $R \to I$) are *guarded* to enforce preconditions. For example, moving from Recognition to Intelligence requires:

- A completed anomaly report with ATT&CK mappings (output of φ_R),
- Provenance for all shared indicators (governance G).

Governance (G) Encodes jurisdictional and bias controls:

- Reciprocity contracts for CTI sharing (R. A. Hagen 2025c),
- Redaction rules for sensitive data,
- Mandatory bias checks (e.g., "Confirm or automation bias" (R. A. Hagen, Øverlier and Helkala 2025)).

Operators ($\Phi = \{\varphi_S\}$) Each phase S is modelled as a *partial operator* $\varphi_S : \mathrm{X}_S \times G \to (\mathrm{Y}_S, \mathrm{A}_S)$, where:

- X_S: Input state (e.g., raw log anomalies in R),
- Y_S: Output state (e.g., prioritized ATT&CK techniques),
- A_S: Audit log .

Example (Synthesis): φ_S takes hypotheses from I, applies human-AI fusion, and logs decisions with bias mitigation rationale.

Metrics (Π) Observable, evidence-grounded measurements:

- Phase latency ,
- Uncertainty reduction (ΔU),
- Action conversion rate.

4.4.2 Invariants: Mathematical Guarantees for SOC Reliability

Invariants are properties that *always hold true* during execution, addressing critical SOC challenges:

Invariant I1 (Progress) Definition: Every enabled transition must either:

- Strictly reduce uncertainty U , or
- Add verifiable information enabling future reductions (e.g., governed CTI in I (R. A. Hagen 2026)).

Why it matters: Prevents "analysis paralysis" (R. A. Hagen, Helkala and Øverlier 2026) by enforcing measurable advancement. The guard $\gamma_{R \to I}$ blocks transition until R outputs actionable hypotheses.

Invariant I2 (Audit Monotonicity) Definition: The log L (concatenation of all A_S) is append-only and tamper-evident. **Why it matters**: Ensures 100% reproducible incident reviews (R. A. Hagen 2025a), solving fragmentation in traditional SOC logs (R. A. Hagen 2026).

Table 6: How Invariants Address SOC Pain Points

Invariant	SOC Challenge	Empirical Link	Outcome
I1 (Progress)	"Endless log reviews"	A1, A6	faster validation (R. A. Hagen 2025d)
I2 (Auditability)	"Who did what, and why?"	A3, A5	Forensically sound trails

4.4.3 End-to-End Guarantees

The composed operator $\Phi* = \phi M \circ \phi S \circ \phi I \circ \phi R \circ \phi P$ processes incidents with three key properties:

1. **Progress (I1)**: No phase transition without uncertainty reduction or verifiable info gain.
2. **Auditability (I2)**: L provides a tamper-evident trail for compliance (R. A. Hagen 2025a).
3. **Measurability**: Π metrics (e.g., ΔU, phase latency) are validated in drills (Section 6).

Comparison to existing frameworks. Unlike MITRE ATT&CK (Strom et al. 2020) or NIST CSF 2.0 (National Institute of Standards and Technology 2024), PRISM-APT's invariants provide:

- **Mathematical guarantees** for progress and auditability,
- **Empirical grounding** in A1–A9's practitioner observations,
- **Operational flexibility** via tool-agnostic operators (φ_S).

Practitioner Summary For SOC teams, invariants ensure:

- **No busywork**: I1 forces hypothesis refinement or info gain at every step.
- **No cover-ups**: I2 makes all actions traceable to people/policies.
- **No vendor lock-in**: $\mathrm{X_S}/\mathrm{Y_S}$ (R. A. Hagen and Helkala 2024; R. Hagen, Øverlier and Helkala 2024).

This formalism bridges theory and practice, offering SOCs the rigor of mathematical models *without sacrificing real-world applicability*.

5. Design Rationale and Traceability

We justify each model element by linking it to one or more items in the derivation map (Table 5), contrasting its role against anchors such as MITRE ATT&CK (Strom et al. 2020) and NIST CSF 2.0 (National Institute of Standards and Technology 2024). This ensures auditability of design choices, support for governed CTI reciprocity, tying the model's phases to empirical evidence (A1–A9) and addressing RQ1–RQ3 holistically.

For Preparation (P) elements like telemetry baselines and script curation (from A9, A8, A7), the rationale stems from fragmented logs and script gaps exacerbating vendor lock-in. We propose standardization to enable behaviour-centric rules, extending ATT&CK mappings beyond static IoCs to portable, tested artefacts, reducing "isolation penalties" for smaller SOCs (A2) and addressing RQ1's framework barriers.

Recognition (R) cues and scripted inquiry derive from anomaly-led needs (A5, A7, A8), where vendor blind spots hinder weak-signal detection. The proposal integrates provenance-aware metrics, contrasting Unified Kill Chain's linearity (Hutchins, Cloppert and Amin 2011) by enabling hypothesis sharpening, filling practical gaps in evolving campaigns (A5).

Intelligence (I) reciprocity contracts (A4, A2) rationalize governed sharing amid regulatory frictions, proposing tiered redaction for trust. This operationalises CTI beyond ENISA overviews (European Union Agency for Cybersecurity (ENISA) 2024), proposing sovereignty-by-operation to mitigate isolation, per RQ3.

Synthesis (S) human-AI fusion and bias mitigation (A1, A6) derive from trust issues and cognitive biases impacting judgments. We propose explainable gates to counter "tunnel vision," advancing NIST's governance by embedding bias controls, ensuring human accountability in AI-driven decisions (RQ2).

Mitigation and Measurement (M) S-CAP metrics and evidence hooks (A3, A5, A2) justify operational measurab-ility, deriving from sovereignty as an "operating state." This proposes drills for auditability, contrasting ISO 27035's principles (*ISO/IEC 27035-1:2023 Information security incident management — Part 1: Principles of incident management* 2023) with quantifiable outcomes, enabling portability dividends.

Overall, these rationales tie PRISM-APT as a minimal, incremental model: evidence-grounded , and behaviour-centric rules over vendor grammars, and governable . By addressing RQ1–RQ3, it bridges theoretical anchors to practice, with invariants (I1–I2) ensuring progress without busywork.

6. Evaluation

6.1 Aim and design.

We evaluate PRISM-APT on three operational properties: (i) ATT&CK technique coverage, (ii) cyclical auditability and provenance of changes, and (iii) sovereignty-by-operation (SoO) as measured via S-CAP thresholds.

Planned methodology. Drills will be run in a controlled SOC-like environment with replayable log sets and scripted operator prompts. We will sample n recent APT campaigns from the programme corpus (A1–A9) using the inclusion gates (state backing, persistence, sophistication), stratified by sector and technique mix. For each campaign, we will execute the PRISM-APT cycle end-to-end with fixed toolchains and pre-registered Π metrics (phase latencies, uncertainty reduction ΔU, action conversion). Pass/fail checks (behaviour coverage, auditability, and SoO via S-CAP) are applied exactly as specified below. No results are reported in this paper; this protocol is included to make the planned evaluation reproducible.

Note. The evaluation described is a protocol for future drills and experiments; no results are reported in this paper.

6.2 Datasets and inclusion gates

We use the programme corpus established in A-studies with the existing inclusion logic based on indicators of state backing, persistence and sophistication.

6.3 Evaluation checks

We define three simple checks. Each check is recorded as pass/fail with a brief note linking to the relevant A-study or artefact lineage.

1. **Behaviour coverage:** for each sampled campaign, at least one behaviour-centric detection (rule or scripted inquiry) is present and executable for every *observed* ATT&CK technique in the sample. *Pass* if all observed techniques have coverage; otherwise *fail*.
2. **Auditability:** the provenance chain is complete for sampled detections, from CTI receipt, through artefact change control and review, to detection event and final disposition, with timestamps and responsible roles. *Pass* if every sampled detection has a complete chain; otherwise *fail*.
3. **SoO via S-CAP:** declared per-phase thresholds (e.g., redundancy, lawful routing, and reciprocity compliance) are met, and the resulting phase profile is recorded. *Pass* if all declared thresholds are satisfied; otherwise *fail*.

Practical acceptance rule. PRISM-APT is considered ready for adoption if it is non-inferior to the baseline on behaviour coverage, satisfies auditability, and meets the declared S-CAP thresholds. Any failed check triggers a targeted revise-and-repeat for the affected phase, using the same conditions.

7. Limitations and Threats to Validity

While PRISM-APT is grounded in A1–A9, limitations include sampling bias in campaigns generalizability across sectors/jurisdictions. Vendor translation fidelity for rule, and assumptions about maturity/logging baselines.

Threats: Qualitative bias in interviews; attribution uncertainty in datasets. These are mitigated by triangulation, but warrant further validation in future work.

8. Conclusion

This paper presents PRISM-APT as a model-first synthesis for governed APT defence, bridging gaps between theoretical frameworks and operational practice. By deriving a cyclical five-phase model from nine empirical studies (A1–A9), we operationalise behaviour-centric rules, reciprocity contracts, and sovereignty-by-operation metrics, empowering resource-constrained organisations to achieve resilient, collaborative defences.

Key contributions: the PRISM-APT model with bias-aware human-in-the-loop gates, an explicit evidence-to-model traceability map, an evaluation plan with metrics for coverage and auditability, and implementation guidance for single-org/consortium deployments, directly address our research questions. RQ1's framework barriers are mitigated through behaviour-centric elements extending anchors like ATT&CK and NIST CSF

2.0. RQ2's practitioner observations inform explainable decision gates and AI fusion, countering cognitive biases. RQ3's digital sovereignty is realized via measurable S-CAP thresholds and conformance-as-code, treating capability as an operating state tested in drills.

PRISM-APT advances APT defence by reducing isolation penalties and enabling collaborative dividends, as evidenced in campaign analyses and thematic insights. While limitations like sampling bias warrant caution, the model's formal guarantees of progress and auditability provide a verifiable foundation.

Future work will expand formalisms, datasets, and case studies, further validating PRISM-APT in diverse SOC environments.

Ethics declaration

This research does not involve new collection of personal data beyond anonymised expert inputs previously cleared under programme studies. No additional ethical clearance is required for this manuscript.

Declaration of generative AI and AI-assisted technologies in the writing process

The author used AI-assisted tools to support language editing and formatting: (i) improving clarity, grammar, and idiom; (ii) reducing dyslexia-related spelling/ordering errors; and (iii) troubleshooting LATEX minutiae

(bibliography keys, table widths, listing frames). All AI-suggested edits were reviewed and approved by the author. Any remaining typos are proudly human.

References

European Union Agency for Cybersecurity (ENISA) (2024). *ENISA Threat Landscape 2024*. Athens, Greece: ENISA. url: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024.

Hagen, R., L. Øverlier and K. Helkala (2024). "The Role of Custom Scripting in APT Incident Response". In: *Proceedings of the 17th Norsk informasjonssikkerhets-konferanse (NISK 2024)*. 3.

Hagen, R. A. (2025a). "A Cloud Abstraction Layer for Cooperative Digital Sovereignty". In: *Proceedings of the 2nd International Conference on Digital Sovereignty (ICDS 2025)*. Oslo, Norway: Springer Nature.

(2025b). "Computational Forensics: The Essential Role of Logs in APT and Advanced Cyberattack Response". In: *Proceedings of the 20th International Conference on Cyber Warfare and Security (ICCWS 2025)*. Vol. 20. 1, pp. 582–591.

(2025c). "Global APT Trends and the Need for Multi-National Coordination". In: *Proceedings of the 24th European Conference on Cyber Warfare and Security (ECCWS 2025)*. Vol. 24. 1, pp. 182–191.

(2025d). "Too Small to Stand Alone: Capability Constraints and Cooperative Digital Sovereignty in APT Defense". In: *Proceedings of the 2nd International Conference on Digital Sovereignty (ICDS 2025)*. Oslo, Norway: Springer Nature.

Hagen, R. A. (2026). "The APT Paradox: Sophisticated Simplicity in Nation-State Cyber Operations (2024–2025)". In: *Proceedings of the 21st International Conference on Cyber Warfare and Security (ICCWS 2026)*. Wilmington, USA: Academic Conferences and Publishing International.

Hagen, R. A. and K. Helkala (2024). "The Complexity of Contemporary Indicators of Compromise". In: *Proceedings of the 23rd European Conference on Cyber Warfare and Security (ECCWS 2024)*. Vol. 23. 1, pp. 147–155.

Hagen, R. A., K. Helkala and L. Øverlier (2026). "Understanding APT Defence Through Expert Eyes: Perceived Needs and Gaps". In: *Secure IT Systems. 30th Nordic Conference, NordSec 2025, Tartu, Estonia, November 12–13, 2025, Proceedings*. Vol. 14781. Lecture Notes in Computer Science. Springer Cham.

Hagen, R. A., L. Øverlier and K. Helkala (2025). "Human Factors in AI-Driven Cybersecurity: Cognitive Biases and Trust Issues". In: *Digital Threats: Research and Practice (DTRAP)* 6.4, pp. 1–20.

Hutchins, E. M., M. J. Cloppert and R. M. Amin (2011). *Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains*. Tech. rep. Lockheed Martin. url: https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf.

ISO/IEC 27035-1:2023 Information security incident management — Part 1: Principles of incident management (2023). Geneva, Switzerland: International Organization for Standardization (ISO). url: https://www.iso.org/standard/78973.html.

National Institute of Standards and Technology (Feb. 2024). *The NIST Cybersecurity Framework (CSF) 2.0*. Tech. rep. NIST CSWP 29. NIST. doi: 10.6028/NIST.CSWP.29. url: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf.

Pols, P. (2017). *The Unified Kill Chain*. Cyber Security Academy, Leiden University. url: https://www.unifiedkillchain.com/assets/The-Unified-Kill-Chain.pdf.

Strom, B. E. et al. (Mar. 2020). *MITRE ATT&CK: Design and Philosophy*. Tech. rep. Approved for Public Release 19-01075-28. McLean, VA: The MITRE Corporation. url: https://attack.mitre.org/docs/ATTACK_Design_and_Philosophy_March_2020.pdf.

Appendix A: Formal Model Details

Model overview

We model the PRISM-APT cycle as a governed transition system

$$M = (S, T, G, \Pi, \Phi),$$

where:

- $S = \{P, R, I, S, M\}$ are the phases Preparation, Recognition, Intelligence, Synthesis, Mitigation.
- $T \subseteq S \times S$ are allowed phase transitions (at minimum (P, R), (R, I), (I, S), (S, M); optional self-loops or back-edges are permitted if guarded).
- G captures governance constraints (obligations, provenance, recording, and access rules).
- Π is a set of measurement functions (e.g., phase latencies, uncertainty reduction, and action conversion).
- $\Phi = \{\varphi_P, \varphi_R, \varphi_I, \varphi_S, \varphi_M\}$ are phase operators. A concrete system state is a tuple

$$\sigma = (s, E, A, \mathrm{L}) \in \Sigma = S \times 2^{\mathrm{E}} \times 2^{\mathrm{A}} \times \mathrm{L},$$

where s is the current phase, E is the evidence store, A are actionable artefacts (rules, scripts, tasks), and L is the append-only audit log. Let $\sigma_0 = (P, \emptyset, \emptyset, \langle\rangle)$ denote the initial state.

Guards and phase operators

For each admissible edge $(s, s') \in T$ we define a guard

$$\gamma_{s \to s'} : 2^{\mathrm{E}} \times 2^{\mathrm{A}} \to \{\top, \bot\},$$

stating when the transition is eligible given the current evidence and actions. Each phase operator

$$\varphi_s : \Sigma \to \Sigma$$

transforms (E, A, L) by applying the phase logic, possibly enriching E, producing or refining A, and appending to L. A single step is written

$$(s, E, A, \mathrm{L}) \; \gamma_{s \to s'} \wedge \varphi_s \; (s', E', A', \mathrm{L}'),$$

meaning the guard holds and the operator yields the successor state with s' as next phase.

Operator interface. Each φ_s satisfies a common interface:

$$\phi_s(\mathrm{E}, \mathrm{A}, \mathrm{L}) = (\mathrm{E}', \mathrm{A}', \mathrm{L} \,\|\, \mathrm{r}_s),$$

where r_s is the phase record appended to L (see A.3), and $(\mathrm{E}', \mathrm{A}')$ are the updated evidence and actions. Guards are total predicates; operators are total functions on Σ.

Governance and audit trail

Governance G induces recording and access obligations. We assume an append-only log

$$\mathrm{L} = \langle r_1, r_2, \ldots, r_k \rangle,$$

where each record has the canonical form

$$r_t = \langle \mathrm{ts}, s, s', \mathrm{id}, H(E_{\mathrm{in}}), H(A_{\mathrm{in}}), H(E_{\mathrm{out}}), H(A_{\mathrm{out}}), \mathrm{actor}, \mathrm{obl} \rangle.$$

Here $H(\cdot)$ is a content hash over a canonical serialisation, actor denotes the accountable agent or process, and obl encodes the governance obligations applied. A simple hash chain $r_t.h = H(r_{t-1} \,\|\, r_t)$ (with $r_0.h$ fixed) makes truncation and in-place edits detectable.

Replay. Given L and the stored digests, a replay function reconstructs, for any t, the inputs consumed and outputs produced in step t, and the guard/operator pair invoked. This suffices to re-derive decisions without re-running the live environment.

End-to-end composition

Let the composed operator be

$$\Phi^* = \varphi_M \circ \varphi_S \circ \varphi_I \circ \varphi_R \circ \varphi_P.$$

On paths that respect T and guards, Φ^* maps (P, E, A, L) to (M, E', A', L') where A' contains mitigation-ready artefacts

Invariants

We restate the two invariants referenced in the main text. Proofs are sketches relying on mild, standard assumptions about guards and operator monotonicity.

Invariant I1 (Progress). *Under the fairness assumption that enabled guards are eventually taken, and with a well-founded progress measure μ that strictly decreases on any non-advancing step, every run either reaches M with $A/=\emptyset$ or halts with an explicit failure record.*

Sketch. Let $\mu : \Sigma \to \mathrm{N}$ measure residual work (e.g., unresolved hypotheses, undecided actions). For any step that does not advance s along a strictly acyclic fragment of T, μ must decrease. Since N is well-founded, infinite descent is impossible. With fairness, enabled advancing guards are eventually taken, so either M is reached with $A/=\emptyset$, or a guard declares irrecoverable failure and the run halts with a terminal record in L.

Invariant I2 (Auditability). *For any finite run,* L *contains a verifiable, tamper-evident account of all state transitions and artefact transformations sufficient to reproduce phase decisions ex post.*

Sketch. By construction, each φ_s appends r_t binding inputs and outputs via content hashes and chaining $r_{t-1}.h$. Any deletion or in-place modification breaks the chain. Since guards and operators are deterministic on (E, A) under G, a replay function, using the digests, re-derives the decisions, satisfying auditability.

Measurement interface Π

We expose a small set of measurement functions to the evaluation protocol:

$$\Pi = \{\tau_s, \Delta U, \kappa\},$$

where τ_s is the latency of phase $s \in S$, ΔU quantifies uncertainty reduction between consecutive phases (e.g., information-theoretic or proxy-based), and κ is the action-conversion rate from synthesised artefacts to executed mitigations. These are observational hooks; the appendix does not report results.

Minimal assumptions

The invariants rely on:

A1. Guards are total and sound: if $\gamma_{s \to s'}(E, A) = \top$ then the preconditions for invoking φ_s toward s' hold.

A2. Operators are total, side-effect confined to $(E, A,$ L), and governance-compliant (they append exactly one r_t per step).

A3. The progress measure μ is well-founded and strictly decreases on any non-advancing step.

A4. Fairness: any persistently enabled advancing guard is eventually taken.

Conformance to pass/fail checks

The evaluation uses three pass/fail checks defined in the main text:

- **Behaviour coverage:** synthesised artefacts cover the behaviour set targeted in the selected campaigns.
- **Auditability:** replay from L reproduces decisions with matching digests.
- **Adoption readiness:** actions in A are deployable with recorded provenance and bounded hand-off latency. These checks are *consumption* constraints on the run; they do not alter the formal semantics above.

Frequent Itemset Mining for Model Data Reduction

Blake Johns and Ryan Benton
The University of South Alabama, Mobile, USA

Johns.blake921@gmail.com
rbenton@southalabama.edu

Abstract: Network intrusion detection systems (NIDS) are in constant battle to come up with new techniques, whether it is a new machine learning model or viewing the data in a new form, to stay ahead of the curve. Recent work has shown that using Deep Learning, coupled with an image-based representation of packet-level data, achieve a high detection rate for a wide number of attacks. Key to the image construction was removing information that can cause misalignment or potentially bias in the Deep Learning model. However, while image-based Deep Learning is known to be good at extracting relevant information from images, it is proposed that using additional preprocessing can result in both improving training and detection time while potentially providing insight into bytes that are frequently occurring in attacks. To this end, this paper proposes the use of Frequent Itemset Mining (FIM), which is a powerful tool that allows the identification of frequent (commonly occurring) itemsets (bytes) from transactional data (bytes in a packet). By identifying the frequent bytes from within the packets of a flow, the feature space can be greatly reduced, which would permit a quicker classification time. This approach is tested on image representations of CICIDS-2017 data where the first *n* packets of a flow are transformed into an image and then fed into a Convolutional Neural Network (CNN) model architecture. Results show that the use of frequent itemsets to identify frequent byte positions of a packet yield comparable results to images that do not use FIM to reduce the number of packet bytes. The feature reduction results in the need for 99% less data, which also reduces the training complexity. We also show better or comparable results to models using the non-reduced packets when reducing the number of packets in a flow (ranging from eight packets to five).

Keywords: Itemset mining, Network intrusion detection, Pattern mining, Feature selection, Data reduction

1. Introduction

As more of our lives connect to a network, securing the transmitted data continues to grow in importance. To protect this data, cyber security efforts need to be put in place. These efforts come in the form of intrusion detection, which alerts a user or organization of nefarious activity.

Network intrusion detection systems (NIDS) sniff packets on the network and identify threats based on some condition (typically heuristic or signature based). The goal is to achieve high classification accuracy while also keeping the time it takes to detection near real time.

The desire to take advantage of developing convolutional neural networks (CNN) has driven an emergence in the application of image classification and computer vision techniques to identify network intrusion. This involves converting network data to images. Numerous conversions have been proposed using grayscale, red-green-blue (RGB), and red-green-blue-alpha (RGBA) image formats (Ghadermazi et al, 2024; Ho et al, 2022; Kim et al, 2020; Toldinas et al, 2021). These works use various models and data and have shown strong results.

This conversion from network data to images changes and increases the feature space complexity. With the large number of features, proper feature selection methods are crucial. Frequent Itemset Mining (FIM) was originally designed for the discovery of common itemsets in market-basket scenarios (Agrawal and Srikant, 1994). Using a *support* measure, FIM can extract common sets of items (features) that are correlated with each other.

In this paper, we use FIM for feature selection of frequent bytes, represented as pixels, in images created from network data. Specifically, we focus on the pixels created by the image conversion proposed in the SPIN-IDS framework (Ghadermazi et al, 2024). This conversion process converts the packets from a unique flow to a single RGB image, where the row of the image corresponds to a specific packet in the flow and the colour corresponds to the direction of the packet (forward or backward). The driving research questions are as follows:

RQ1 - Can frequent itemset mining be used to reduce the number of packets required to make an accurate detection using the SPIN-IDS framework?

RQ2 - Can frequent itemset mining be used to identify a subset of the original data that achieves similar results to the SPIN-IDS experimental results by the ninth packet?

We propose a novel feature extraction method utilizing itemset mining. This feature extraction method takes an image representing the packets within a flow and identifies frequently occurring byte patterns by converting the image to a transactional data set.

The remainder of the paper is organized as follows. Section 2 discusses related work and Section 3 outlines the experiments. Section 4 presents the results of the experiments. Section 5 identifies areas for future research and Section 6 presents the conclusions.

2. Related Works

2.1 Image Transformations in Network Intrusion Detection

Network data comes in two forms: raw packets (bytes) and flows. The raw packets include the required information that is dictated by layer 3 of the OSI model (Kumar et al, 2014). A flow consists of the meta-data of a conversation between two endpoints. A single flow is identified by a 5-tuple consisting of source IP, destination IP, source port, destination port, and protocol (Claise, 2008).

Transforming network data to images has been done for both flows and packets. SPIN-IDS takes advantage of the unique tuple to identify the flow but works directly with the packets within that flow to create an image. SPIN-IDS uses the first nine packets in a flow (Ghadermazi et al, 2024). To convert the packet into the RGB (red, green, blue) image format, the Ethernet packet header is discarded along with specific bytes and all flags from the IP and TCP headers. The remaining bytes of the packet, totalling 1486, are converted to hexadecimal byte representation and undergo a decimal conversion (bound by 255) before being put into a vector. The direction of the packet determines the channel to which the data are assigned. Forward packets go to the red channel, backward packets go to the green channel, and the blue channel remains unused. Finally, these packet vectors are combined to create a single image for the flow, at most 9x1486 (height x width) in size.

In a related approach, flow data is converted to both RGB and grayscale images (Kim et al, 2020). This transformation converts non-numeric and symbolic values to numerical values. One-hot encoding is applied to features, such as the protocol, for conversion and all the features are stored within a vector. The feature vectors are joined together to create a single vector with values bound by [0, 255] and reshaped to fit the image space of size (13 x 9) pixels. The new data are fed into 3 channel and single channel image representations for RGB and grayscale images. Additionally, network flows have been converted to RGBA images through numerical conversions (Toldinas et al, 2021). Specifically, the values are converted to unsigned 32-bit integers; if the value is a float, then the remainder is truncated. These values are then mapped to a byte array, where every array is 168 bytes, and every 4 bytes represents a pixel in the ARGB (alpha, red, green, blue) image. This conversion outputs images of size (42 x 1) and a size of (7 x 6).

Rather than using an entire flow, a time window is specified to generate new features with respect to the elapsed time of the window which are converted into an image (Ho et al, 2022). This elapsed window is applied to all flows occurring within that window. It then uses a decision tree for feature selection, the Mahalanobis Distance-Based Oversampling (Abdi and Hashemi, 2016) for class balancing and identifying features with an information gain of greater than 0.001 in at least one attack type. Through this process, 24 features were identified for use in the image conversion process.

Another work uses the DeepInsight Python tool (Alok Al Lab, 2025), a Kernel PCA (KPCA) (Scholkopf et al, 1998), and a Gabor filter to convert network flow data into an image using a suitable user-defined image size (Siddiqi, Ahmed, and Pak, 2022). Missing values and duplicate samples are removed from the data. Non-numerical features were then masked using one-hot encoding before being passed into DeepInsight. Next, KPCA converts the data into a 2-dimensional space. Using an appropriate image size, a Gabor filter is then applied to obtain the data in frequency space while also keeping respect to time space. The image size varies between data sets and is (9 x 9) for the CSE-CIC- IDS 2017 / 2018 and ISCX-IDS 2012 while the NSL-KDD data set uses a 12 x 12 size image. The data is normalized, resulting in a 2-dimensional grayscale image.

2.2 Itemset Mining

An item or itemset comes from the notion of basket data, where the items within the basket represent a transaction and those collections collectively form transactions over time. Frequent itemset mining identifies items that frequently occur together within databases and forms relationships between these items, known as association rules (Agrawal et al, 1993). A persistent itemset mining issue is how memory intensive it is when presented with large amounts of data, particularly many features or a lower support threshold. This was shown in comparisons of the FP-Growth (Han et al, 2004) and Eclat algorithms (Zaki, 2000) followed by the

Eclat and Medic algorithms (Goethals, 2004). A lower support generates more item sets and states that if a database cannot be completely loaded into main memory, then memory problems will arise when generating item sets (Goethals, 2004).

Numerous works have been performed to reduce the memory requirement of itemset mining tasks. The introduction of the CFP-Tree and CFP-array applied lightweight compression techniques to reduce the memory footprint (Schlegel et al, 2011). This compression resulted in a decrease in memory usage of 7x - 25x. Influenced from machine learning, a test-train split method is used to split the data into different partitions (Al-Zeiadi, Mohammed, and Al-Maqaleh, 2025). Refraining from loading all the data into memory allows this method to efficiently handle large datasets.

Itemset mining is traditionally used with transactional data; however, some work attempts to apply this to domains such as image classification. To obtain the image in the proper representation to perform FIM, the image is transformed into a set of visual words (Nowozin et al, 2007). A backtracking FIM approach is applied to these words to extract the frequent itemsets. Then a support vector machine (SVM) is trained on the item sets.

Another improvement uses a custom itemset mining algorithm to create a grouplet, which is structured information of an image (Yao, Bangpeng, and Fei-Fei, 2010). Once the initial grouplets are created, itemset mining is applied to them to prune out infrequent grouplets. The mining of different configurations within images is also proposed (Quack et al, 2007). These configurations are used to separate background information from the primary image and allow focus on the object of interest. Fernando et al (2012) defines an itemset as a pair that contains a visual word and its frequency (w, s). The set of transactions are created from the set of local histograms (local bag-of-words) computed from the image. Frequent Local Histograms (FLH) are then formed, and any frequent mining algorithm can be used to extract patterns which are used directly for image classification.

3. Experiments

The experiments require the generation of images that represent the packets within a network flow. These RGB images are at most 9 pixels in height. The height of the image depends on the number of packets selected as a cap for that flow. These experiments begin with the first nine packets in a flow and continue to reduce this down to the first five.

The design of the experiments is two parts. Firstly, packet bytes are extracted from the RGB image and converted into a transactional representation. This transactional representation treats bytes in a single row as a set of items, allowing the application of itemset mining. After the transformation, the transactions are placed into two datasets. The first data set is the wholistic transactional dataset, which contains every row of pixels from each image. The second data set is created with respect to the position of the row within the image. As each row in the image corresponds to a packet in a flow, this data set contains positionally relative data within the image. Once the transactional data sets have been created, frequent itemsets (sets of items) will be extracted. The criteria for an itemset to be considered frequent is a measure of support, defined as *count(x)/count(dataset) >= MinSupport* where x is an item in the database and *MinSupport* is a decimal value threshold that represents a percentage. This metric is relative to the size of the data and is expected to yield different results between the full transactional data set and the packet position data sets.

Once the frequent itemsets have been extracted, the byte positions within the itemsets will be used as the new feature space. The original data set begins with a maximum size of (9, 1486, 3) representing (packet, bytes, channel). This is a maximum size because not all images will have a height of nine, but it remains the cap. Assuming that there were 10 itemsets extracted that contained eight frequent byte positions, the new size of the feature space is (9, 8, 3).

3.1 Experimental Setup

The experiments were run on a ROG Zephyrus G14 laptop with the Windows operating system. This machine has 16 GB of RAM and an AMD Ryzen 9 6900HS processor. The experiments used the CICIDS 2017 data set (Sharafaldin, Lashkari, and Ghorbani, 2018). Network traffic is converted to images using a Python implementation of the image generation and byte selection method defined by the SPIN-IDS framework (Ghadermazi et al, 2024). Further data transformations and itemset mining are also implemented in Python. The CNN model was built using PyTorch (Paszke et al, 2019).

3.2 Dataset

The CICIDS 2017 data set consists of 13 attacks that cover denial of service (DoS), brute force, web attacks, infiltration, and reconnaissance activities. For these experiments, the network data have been reordered using reordercap and duplicate packets have been removed using editcap; both applications are a part of the Wireshark distribution. The labelling effort is strictly with the IP address of the attacker's machine (Pre-NAT: 205.174.165.73, Post-Nat: 172.16.0.1) after going through the firewall and follows a binary scheme of 0 (benign), 1 (threat). This labelling effort is likely to affect the results of the modelling and itemsets, as many works in the cyber security community that have identified disparities in both the attacks that were effective and the time of which the attack is being executed (Liu et al, 2022).

4. Image Generation

To create the images, the algorithm proposed by the SPIN-IDS framework was used. The implementation was done using Python 3.12 and uses the Python module *scapy* (The Scapy Community, 2025) for packet dissection targeting TCP packets.

Each image represents a single TCP flow and is identified by the flow tuple (source IP, destination IP, source port, destination port, and protocol). A cap for the number of packets in each flow is set to 9 and reduced by 1 until the cap is set to 5. Each time the cap is reduced, a new set of images is created.

Within the packet, the Ethernet header is discarded. We continue to follow the process described by the SPIN-IDS framework and select bytes 3-9 and 11-12 from the IP header, bytes 5-20 of the Transport header, and the entirety of the payload to create a packet vector. These byte values are then converted to decimal equivalent of their hex value and are bound from [0, 255]. If the total number of bytes selected exceeds 1486 then the remainder is discarded. If the number of bytes is less than 1486 then the vector is padded with 0's to ensure that there is uniform shape.

4.1 Image Transformation

This section details the process of transforming images into transactional data sets. This representation is in the form of a CSV where each item separated by a comma is associated with the channel, the byte position, and the value in that byte position. As an example, the packet vector Red Channel [0, 25, 0, 250] would be reduced to [REDb1v32, REDb3v256].

The packet vector [[0, 25, 0, 250], [0, 0, 0, 0], [0, 0, 0, 0]] is read in and flattened to become [0, 0, 0, 25, 0, 0, 0, 0, 0, 250, 0, 0, 0, 0, 0]. Once flattened, the starting indices 0 and 1 are set for the red and green channels, respectively, and every third value from the starting positions is taken to separate the channels. To identify which channel is active in the packet vector, the mean of the values is taken and the channel whose mean is greater than zero is the active channel. The blue channel is disregarded, as this channel is only used for image shape and any non-zero value in this channel indicates that there was bleed over.

Once the channel has been identified, 0's are removed. This removal is done because of an added padding that is done during image creation. The original index positions are stored to avoid losing that information during zero removal. The first arrow in Figure 1 shows the before and after of the transformation. At this stage, the possibly values for each index are 0-255. To combat feature complexity, we employ a binning strategy.

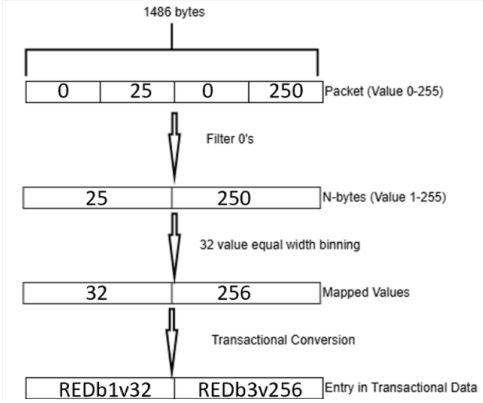


Figure 1: Data Transformation

Equal-width binning is used to reduce the number of possible values from 256 possible values to eight. Each bin covers 32 values, resulting in the eight bins. The second arrow in Figure 1 shows the value 25 in index 0 mapped to the 32 bin and the value 250 in index 1 mapped to the 256 bin.

Finally, the values are converted to their transactional form, which consists of the channel associated with the row in the image, the initial index (byte position), and the binned value. For example, Figure 1 shows 32 mapped to REDb1v32 and 256 mapped to REDb3v250. Following the trace back, the original channel was red, the original index was 1 (despite being in index 0 after the final transformation), and the binned value is 32. With this transformation, the previous vector Red Channel [0, 25, 0, 250] is now a transaction {REDb1v32, REDb3v256}.

For each image, two data sets are created with respect to the data label. The first data set is the wholistic transactional data set. In this dataset, each entry is associated with a transformed packet from a flow. This data set includes every packet of every flow as a complete record of transactions.

The second data set created is focused on the position of the packet (row of an image) within each flow corresponding to a label. For example, if a flow consisted of 4 packets, then each transformed packet vector would go to a separate file: packet 1: packet 1 transactions, packet 2: packet 2 transactions, etc. This is done for every image and generates a single, positionally based data set for the packets of the flows of a label.

4.2 Frequent Itemset Extraction

To extract frequent itemsets we use the Python module mlxtend. This module supplies a priori approach, priori stating that if an itemset is frequent, then all subsets must also be frequent. To determine what itemsets are frequent, the minimum support is initially set to 0.4 (40%).

4.3 Modelling

A CNN is used for the classification of image data. The model is implemented as specified by the SPIN-IDS framework to ensure a likewise comparison. This framework uses four convolution layers, three max pooling, batch normalization, two instances of dropout, and two dense layers. The model implementation is performed using the PyTorch deep learning library in Python and is evaluated using the accuracy metric.

5. Results

5.1 Itemset Extraction

Itemsets were extracted from all data sets using a minimum support of 0.4. Looking at the itemsets present in more than 40% of the data in each transactional database, we can identify four key bytes for potential classification: byte {2, 5, 7, 19} rather than the entire 1486 bytes included in the image.

Analysis of packet 1 data sets shows that the highest itemset support within the benign label is 0.99 and 1.0 within the threat label. Looking at itemsets with support greater than 0.8 support, the itemsets are {REDb19v128, REDb2v64, REDb5v64} with a support of 0.93 occurring together within the benign label. The threat label for packet 1 consist of longer itemsets such as {REDb2v64, REDb19v32, REDb5v64, REDb7v64, REDb18v160, REDb21v32, REDb20v128} with the highest singleton itemset (itemset of size 1) support being {REDb5v64} as 1.0 or 100% of the time occurring in a threat packet. This analysis is applied to every set of packet transactional data.

Through itemset mining, we can reduce the dataset from 9x1486x3 to 9x26x3. This new feature space contains the byte positions that were frequent in both threat and benign labels, as well as the entire transactional data set and the packet positional data sets. The feature space extracted is as follows: byte positions {0, 2, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 30}.

5.2 Model Performance

To adequately answer the research questions, we evaluate the model by examining overall accuracy. Overall accuracy is calculated by (TP +TN) / (P + N) where TP is true positive, TN is true negative, total positive P, and total negative N.

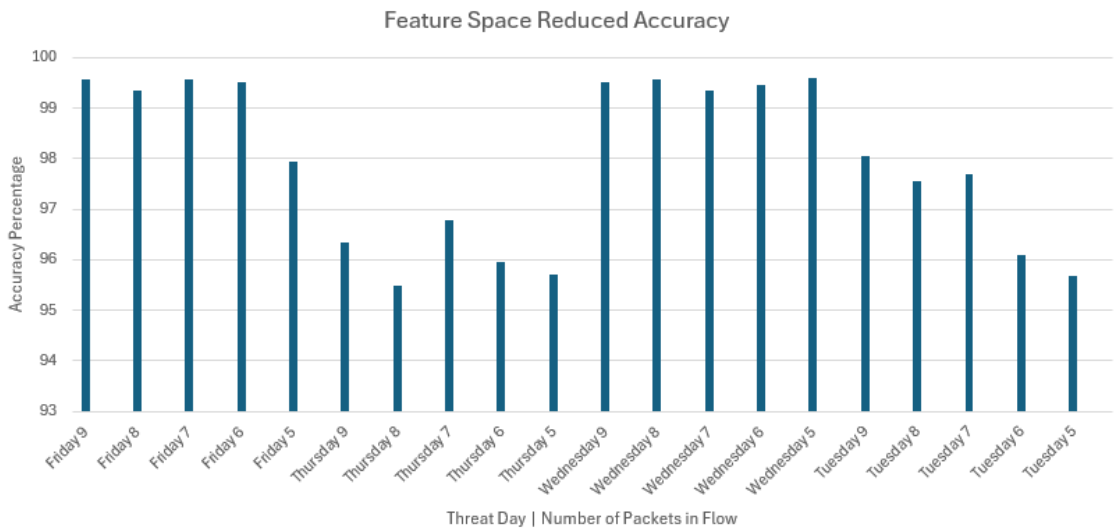


Figure 2: Frequent Itemset Reduced Feature Space Model Accuracy

For each day, the model was trained separately from the others. In doing so, an accuracy of more than 95% was achieved on all days as seen in Figure 2. Furthermore, the amount of time required to train the model with the reduced feature space is noticeably shorter than the model without a reduced feature space as seen in Figure 3.

With these results, initial research questions can be answered. To answer RQ1, we applied the SPIN-IDS framework using a CNN that has been trained and tested with the reduced dataset. Rather than 1,486 features, this CNN is trained using only the 23 identified features. Furthermore, there are four variations of the model trained, each trained on a different number of packets from within the flow (ranging from the first five packets through the first eight packets). All models will use a 70:30 train / test split.

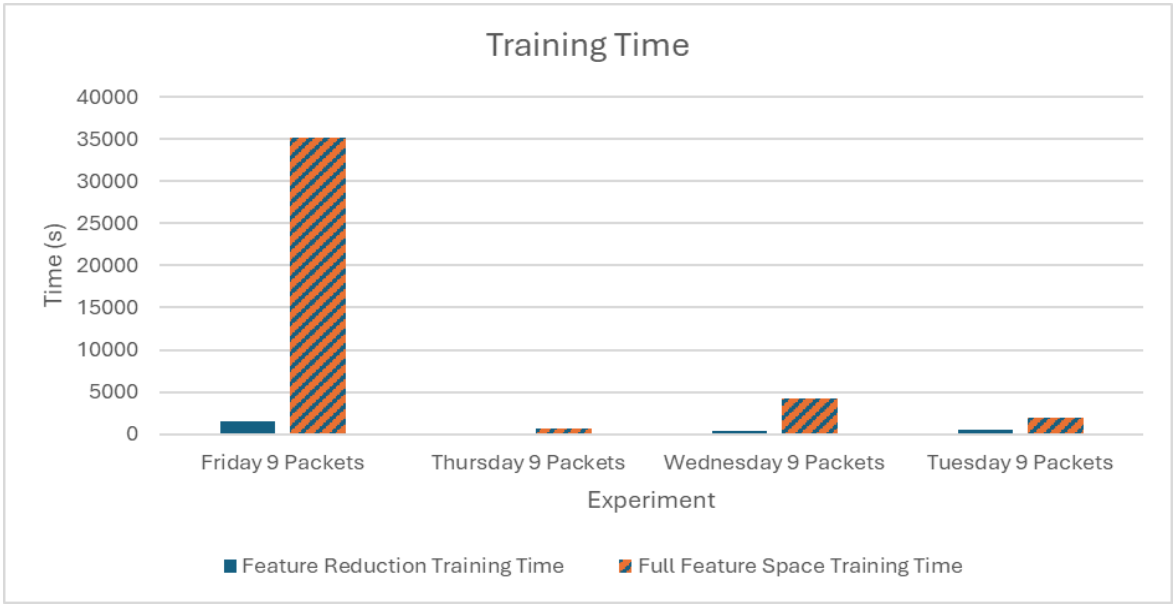


Figure 3: Training Time Comparison between Reduced Feature Space and Full Feature Space

With the reduced data set, the model has shown results of 99% accuracy on Friday with 8 packets (continuing to maintain an accuracy of above 96% down through the fifth packet). Thursday varied on the results, likely due to previously mentioned issues with the dataset but varied between 95-96% accuracy across all runs. Wednesday experiments maintained an accuracy of 99% throughout the reduced number of packets and Thursday saw a gradual decline from 98% to 95% going from nine packets to five.

Given that the SPIN-IDS framework reports 98% overall accuracy across all days, using this reduced data set achieves comparable results at earlier packets within a flow.

Answering RQ2 requires that the CNN model be trained for a binary classification task rather than the author's original multi-class classification. Rerunning the experiments set to answer RQ1 with the full feature space allow us to answer RQ2. In most experiments, the model using the reduced feature space either outperforms the model with the full feature space or has comparable results (differing by 0.5% accuracy). These results can be seen in Figure 4. There are experiments where the model using the full feature space never converged, specifically Thursday 9, Thursday 7, Wednesday 9, Wednesday 6, and Tuesday 6. Modifications to these model runs were not done as to keep experimental processes as closely comparable as possible.

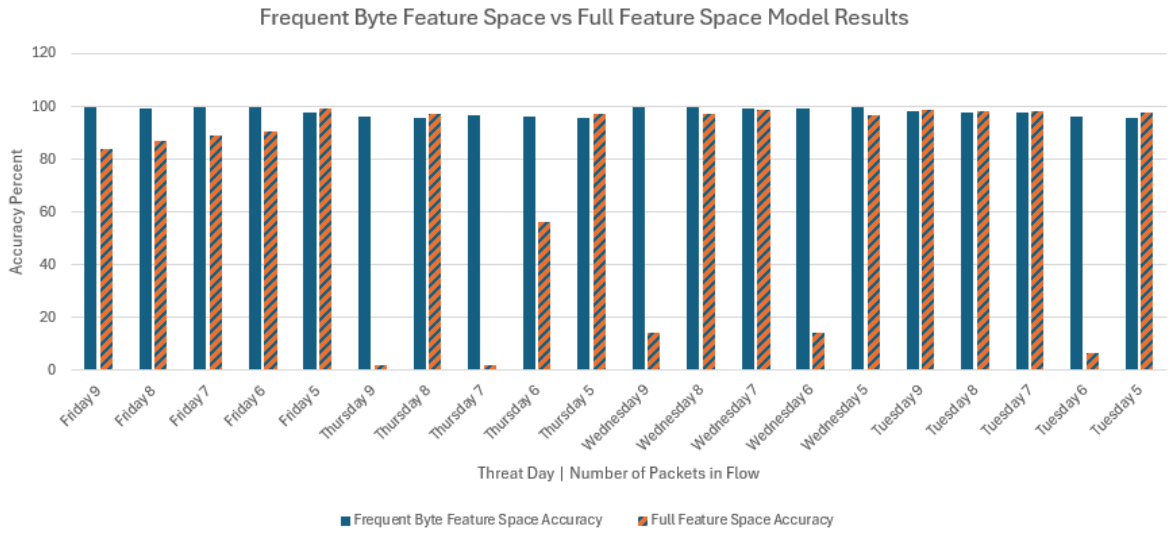


Figure 4: Comparing Model with Reduced Feature Set to Full Feature Set

6. Areas of Future Research

The data transformation efforts proposed here have given the ability to view the data in the form of transactions. In addition, the data split into subsets based on the position of the packet within the flow can be used to consider sequential rule mining. In this sense, frequent itemsets can be extracted from each packet positional dataset and merged to create a set of ordered sequential occurrences with respect to the associated bytes. This makes more sense given that the flows are based on the protocol that two endpoints are talking across, and if there is common behaviour among that protocol during conversation, then these rules could bring light to it at the byte level.

A limitation of this work is that we only used the CICIDS-2017 dataset. While the results here are promising, future work should include testing against other datasets while evaluating the impact of the feature selection method against real-time performance.

7. Conclusions

In this work, we have proposed a novel data transformation approach and application of itemset mining for the extraction of frequent packet bytes represented as images. This effort has brought forth a new way to view network data, as a transactional data set. Through this transformation, it has successfully been shown that frequent itemsets of various complexity exist not only in the wholistic transactional data set but also with respect to the packet position within the flow itself.

These itemsets can be used to reduce the feature space for image classification of image-based network intrusion detection. With results comparable to other CNNs, the proposed data transformation and feature selection method uses 1.5% of the original data (reducing the feature space from (9, 1486, 3) to (9, 23, 3)). With the reduction of data, noise is also filtered out, and accuracy is sustained at less packets within the flow (testing from the first nine packets to the first five packets in a flow).

Ethics declaration: No ethical clearance was required for the research conducted in this paper.

AI Declaration: AI tools were not used.

References

Abdi, L. and Hashemi, S., 2015. To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE transactions on Knowledge and Data Engineering*, *28*(1), pp.238-251.

Agrawal, R., Imieliński, T. and Swami, A., 1993, June. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (pp. 207-216).

Agrawal, R.S. and Srikant, R., 1994, September. R. Fast algorithms for mining association rules. In *Proceedings of the 20th International Conference on Very Large Data Bases, VLDB* (pp. 487-499).

Al-Zeiadi, M.A. and AI-Maqaleh, B.M., 2025. Incremental Closed Frequent Itemsets Mining based Approach Using Maximal Candidates. *IEEE Access*.

Alok AI Lab (2025) pyDeepInsight. Available at: https://github.com/alok-ai-lab/pyDeepInsight (Accessed: October 2, 2025)

Claise, B., 2008. *Specification of the IP flow information export (IPFIX) protocol for the exchange of IP traffic flow information* (No. rfc5101).

Fernando, B., Fromont, E. and Tuytelaars, T., 2012, October. Effective use of frequent itemset mining for image classification. In *European conference on computer vision* (pp. 214-227). Berlin, Heidelberg: Springer Berlin Heidelberg.

Ghadermazi, J., Shah, A. and Bastian, N.D., 2024. Towards real-time network intrusion detection with image-based sequential packets representation. *IEEE Transactions on Big Data*, *11*(1), pp.157-173.

Goethals, B., 2004, March. Memory issues in frequent itemset mining. In *Proceedings of the 2004 ACM symposium on Applied computing* (pp. 530-534).

Han, J., Pei, J., Yin, Y. and Mao, R., 2004. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, *8*(1), pp.53-87.

Ho, C.M.K., Yow, K.C., Zhu, Z. and Aravamuthan, S., 2022. Network intrusion detection via flow-to-image conversion and vision transformer classification. *IEEE Access*, *10*, pp.97780-97793.

Kim, J., Kim, J., Kim, H., Shim, M. and Choi, E., 2020. CNN-based network intrusion detection against denial-of-service attacks. *Electronics*, *9*(6), p.916.

Kumar, S., Dalal, S. and Dixit, V., 2014. The osi model: overview on the seven layers of computer networks. *International Journal of Computer Science and Information Technology Research*, *2*(3), pp.461-466.

Liu, L., Engelen, G., Lynar, T., Essam, D. and Joosen, W., 2022, October. Error prevalence in nids datasets: A case study on cic-ids-2017 and cse-cic-ids-2018. In *2022 IEEE Conference on Communications and Network Security (CNS)* (pp. 254-262). IEEE.

Nowozin, S., Tsuda, K., Uno, T., Kudo, T. and BakIr, G., 2007, June. Weighted substructure mining for image analysis. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, *32*.

Quack, T., Ferrari, V., Leibe, B. and Van Gool, L., 2007, October. Efficient mining of frequent and distinctive feature configurations. In *2007 IEEE 11th International Conference on Computer Vision* (pp. 1-8). IEEE.

Siddiqi, M.A. and Pak, W., 2022. An Optimized and Hybrid Framework for Image Processing Based Network Intrusion Detection System. *Computers, Materials & Continua*, *73*(2).

Schlegel, B., Gemulla, R. and Lehner, W., 2011, March. Memory-efficient frequent-itemset mining. In *Proceedings of the 14th international conference on extending database technology* (pp. 461-472).

Schölkopf, B., Smola, A. and Müller, K.R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5), pp.1299-1319.

Sharafaldin, I., Lashkari, A.H. and Ghorbani, A.A., 2018. *Toward generating a new intrusion detection dataset and intrusion traffic characterization*. *ICISSp*, *1*(2018), pp.108-116.

The Scapy Community (2025) Scapy. Available at: https://scapy.net/ (Accessed: October 2, 2025)

Toldinas, J., Venčkauskas, A., Damaševičius, R., Grigaliūnas, Š., Morkevičius, N. and Baranauskas, E., 2021. *A novel approach for network intrusion detection using multistage deep learning image recognition*. *Electronics*, *10*(15), p.1854.

Yao, B. and Fei-Fei, L., 2010, June. Grouplet: A structured image representation for recognizing human and object interactions. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 9-16). IEEE.

Zaki, M.J., 2002. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, *12*(3), pp.372-390.

Sleeper Code in Protein Data Files as Cyber Adversarial Vectors

Tia Pope
North Carolina A&T State University, Greensboro, North Carolina, USA

tpope@aggies.ncat.edu

Abstract: Scientific protein data formats are widely assumed to be inert, yet their use in automated and AI-driven research environments creates overlooked pathways for cyberbiosecurity risk. This work examines sleeper code patterns, defined as structured non-executable strings embedded in FASTA files, CIF metadata, and protein sequences that persist through downstream processing. A controlled simulation framework models ten representative adversarial use cases and reveals that many workflow components carry these patterns forward without removal, including cloud alignment tools, high-performance computing pipelines, visualization utilities, and transformer-based protein models. Across the ten simulations, nine workflows preserved at least one embedded pattern, which confirms broad systemic tolerance for structured symbolic content. Results show that permissive parsing rules and AI prefix conditioning allow symbolic content to survive reformatting and, in some cases, to become further embedded within generated outputs. These findings indicate a structural blind spot in scientific workflows where biological trust assumptions obscure computational vulnerabilities. To address this gap, the paper introduces a multilayer mitigation framework that combines input sanitation, anomaly detection, AI model guardrails, workflow provenance, and federated containment. Taken together, the study reframes protein data formats as potential cyber vectors and highlights the need for interdisciplinary approaches that strengthen digital resilience across computational biology and national research infrastructure.

Keywords: Cyberbiosecurity, Data formats, Data integrity, Adversarial data, FASTA, mmCIF, AI protein models

1. Introduction

Digital transformation has reshaped modern cyber conflict and expanded attack surfaces across scientific and biomedical domains (Murch *et al.*, 2020; National Institute of Standards and Technology, 2022; National Science Advisory Board for Biosecurity, 2023). These domains support national research capacity and pharmaceutical development. They also enable strategic biotechnology innovation. As computational biology and artificial intelligence become central to scientific practice, the integrity of biological data has become inseparable from the integrity of digital systems (National Institute of Standards and Technology, 2022; Urbina et al., 2022). Despite this convergence, the cyberbiosecurity risks within scientific data formats remain overlooked.

In this paper, cyberbiosecurity is defined as the discipline that protects the confidentiality and integrity of biological data and computational tools from cyber threats. The term also includes the protection of models and workflows that operate across the digital biological interface. This definition reflects a landscape in which biological information is a scientific resource and a digital asset. It is therefore vulnerable to cyber manipulation and adversarial misuse.

Protein data files, such as FASTA, PDB, or CIF, are often treated as neutral scientific objects. They move through cloud services and bioinformatics pipelines with little sanitization (Ney et al., 2017; wwPDB Consortium, 2019). They also enter laboratory information systems and high performance computing (HPC) clusters with limited verification. Many of these systems rely on trust shaped by biological reasoning rather than computational security practice. This reliance creates hidden attack surfaces within the data that support modern biotechnology (Gore et al., 2017).

Advances in AI-driven protein modeling introduce new complexity. Large sequence models accept many forms of amino acid input and can generate extended sequences that feed into engineering workflows (Carbone et al., 2022; Alkhouri et al., 2020). These systems may preserve structured patterns that remain invisible to human review. Such patterns are not executable but they can persist across storage and computation. Their quiet movement through scientific systems reveals a neglected path for cyber exploitation.

This gap motivates the central question of this work. **Can protein data formats serve as covert adversarial vectors, enabling the silent spread of structured patterns across scientific and AI-driven workflows?**

Modern protein science relies on open, collaborative workflows. Data files routinely pass between contributors, repositories, and automated processing systems. The assumed adversary is a legitimate participant in this pipeline, such as a contributor, collaborator, or upstream repository source. No software vulnerabilities, system-level access, or code execution capabilities are assumed. Risk arises from permissive handling of standard protein formats. Symbolic content placed in headers, metadata fields, or sequence

regions may persist through routine processing. The concern is persistence rather than execution, allowing structured patterns to propagate across tools and be amplified by AI-driven models. Plausible impacts include contamination of downstream datasets, distortion of provenance records, unintended information channels in federated environments, and erosion of trust in shared biological infrastructure.

To explore this question, we examine the idea of sleeper code in protein data files. Sleeper code refers to synthetic markers that remain dormant during ingestion yet endure many downstream transformations. We simulate ten adversarial use cases to observe how these markers move through pipelines and visualization tools. We also examine their movement through AI models and federated data repositories. These simulations reveal conditions under which scientific workflows may preserve patterns that seem harmless.

Our analysis rests on a simple hypothesis. If scientific workflows treat protein data as safe by default, then structured non-biological patterns will persist across multiple computational stages. This persistence would show that protein files can act as covert channels for adversarial influence. The findings presented here test this hypothesis.

2. Related Work

Cyberbiosecurity has grown as a multidisciplinary field at the intersection of cybersecurity and the life sciences. Early work by Peccoud and Murch defined cyberbiosecurity as an effort to protect the bioeconomy and biotechnology systems (Murch et al., 2020; Peccoud et al., 2018). Later analyses expanded this view to risks in laboratory automation and genomic databases (Murch et al., 2020). They also considered threats to networked instruments and digital design workflows. These studies emphasize the need to secure biological information systems but they focus on infrastructure rather than the structure of scientific data.

A second line of research has examined weaknesses in bioinformatics software and related pipelines. Ney and colleagues showed that unsafe parsing in DNA sequencing tools can enable remote code execution through crafted inputs (Ney et al., 2017). Broader work in digital biosecurity highlights risks in software that supports biological research (National Institute of Standards and Technology, 2022). The work also stresses the need for secure development practices in genomics and related domains. These studies reveal the fragility of biological computation, but they center on software flaws rather than data encoding.

Concerns about biological data as a cyber attack vector appear in several studies. A previously mentioned work also has shown that unsafe parsing of biological inputs can enable unexpected behavior in sequencing software, including buffer overflows triggered by synthetic DNA (Ney et al., 2017). Reviews in cyberbiosecurity note that DNA and genomic data can be manipulated for adversarial purposes (Murch et al., 2020). This includes embedding harmful content in biological samples or sequence files. These studies focus on DNA and genomic systems and do not examine digital protein structure files or AI-driven structural workflows.

Adversarial machine learning work offers further insight. Goodfellow and Papernot showed that small perturbations can reliably alter deep learning outputs (Goodfellow et al., 2015). Later studies applied these ideas to molecular design. Another work crafted adversarial attacks on protein language models that altered predicted properties (Carbone et al., 2022). Alkhouri and colleagues showed that minor sequence changes can shift predicted three dimensional structures produced by AlphaFold (Alkhouri et al., 2020). These studies focus on model behavior rather than the role of scientific file formats as persistent carriers of structured input.

Structural biology research focuses on scientific quality rather than security. The Worldwide Protein Data Bank provides detailed specifications for PDB and mmCIF formats with an emphasis on accurate representation (wwPDB Consortium, 2019). Validation tools such as the wwPDB pipeline and MolProbity aim to improve geometric and experimental correctness (Gore et al., 2017). These systems target scientific reliability. They do not consider adversarial misuse of structural files.

AI-enabled biological design has raised new biosafety concerns. Prior work has demonstrated that molecular design models can be tuned to generate toxic compounds (Urbina et al., 2022). Additional analyses emphasize the need for governance and oversight of biological AI models to mitigate misuse (American Association for the Advancement of Science, 2024). This body of work frames model misuse as a dual use risk. It does not examine cyber pathways that arise from file formats in research pipelines.

Recent U.S. policies underscore the need to secure digital components of the bioeconomy. NIST documents highlight the unique risks associated with genomic data (National Institute of Standards and Technology, 2022). The National Cybersecurity Center of Excellence offers guidance for genomic environments (National Cybersecurity Center of Excellence, 2023). Executive orders call for the protection of biological data from

digital intrusion (White House, 2022; White House, 2025). Additional guidance from the National Science Advisory Board for Biosecurity and the Department of Defense stresses the need to manage cyber risks from emerging computational tools (National Science Advisory Board for Biosecurity, 2023; Department of Defense, 2025). These policies recognize digital biological interfaces as growing cyber risk surfaces but do not address protein structure formats.

Across these bodies of work, a consistent pattern emerges. Research has focused on software vulnerabilities and infrastructure risks, as well as the robustness of machine learning systems. No prior study has examined protein structure files as potential cyber vectors. The closest related efforts appear in our own work, including our analyses published in the 22nd Annual International Conference on Privacy, Security, and Trust conference proceedings (Pope and Patooghy, 2025a) and in the NeurIPS BioSafe GenAI Workshop (Pope and Patooghy, 2025b). These studies address adversarial risks at the biological and computational boundary, yet neither evaluates PDB or mmCIF files as carriers of structured patterns within protein workflows. This work fills that gap by treating protein data formats as possible channels for adversarial influence.

While related to data poisoning and adversarial machine learning, this work addresses a distinct class of risk. Data poisoning focuses on statistical influence over model behavior, whereas sleeper code persistence operates at the data representation level and does not require training-time access or execution. Injection-style attacks rely on software interpretation of malicious input, which this study explicitly avoids. Provenance frameworks track data lineage but typically assume semantic validity of file contents. In contrast, this work demonstrates that structurally valid protein files can act as passive carriers of symbolic patterns that persist across tools, workflow and AI systems. Our framing positions protein data formats themselves as a cyberbiosecurity concern rather than merely a substrate for other attacks.

3. Methodology

We used a controlled simulation framework that examines how protein data formats may preserve non-biological patterns across scientific and AI-enabled workflows. The simulation reflects behavior observed in real tools while removing sensitive operational detail. All markers are synthetic and non-executable and all inputs follow safe patterns. The design captures data transformation rather than system vulnerability. This supports open science and responsible communication.

The simulation executes ten experiments. These experiments represent different adversarial conditions that may appear in scientific workflows. Each experiment uses the same workflow and varies only by marker placement or format location (see **Section 3.9**).

3.1 Research Design Overview

The study examines whether scientific workflows preserve structured patterns in routine protein data fields. The workflow contains ingestion and database interaction, alignment preparation and HPC pipelines (**Figure 1**). It also contains visualization and federated storage stages. Each stage applies a controlled transformation that reflects common practice in protein science. The workflow records marker presence after each stage.

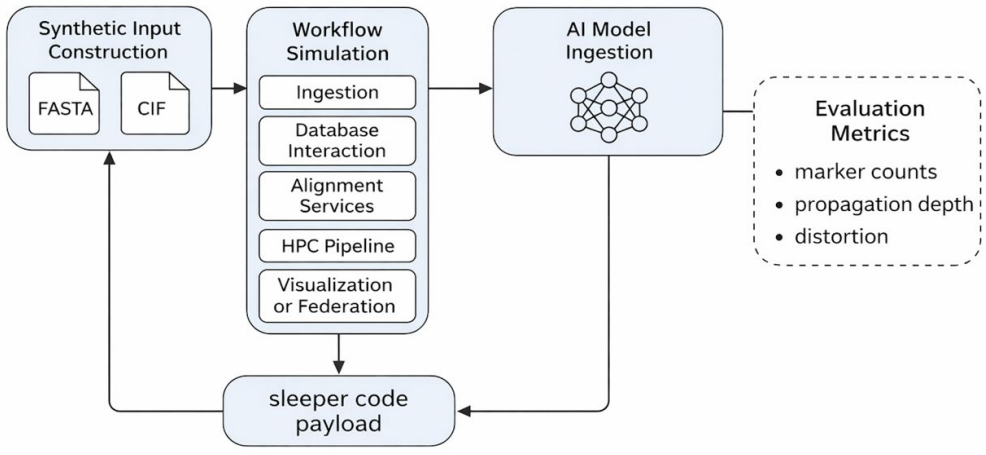


Figure 1: Overview of the simulation workflow. The figure shows how synthetic protein data pass through modelled, workflow stages and AI ingestion to evaluate the persistence and transformation of structured sleeper code markers

This design isolates the behavior of the data itself. It reflects patterns seen in real experiments that we conducted with mainstream tools. It simulates those conditions because direct description would reveal methods that could enable exploitation. It focuses on passive propagation rather than active compromise. It supports a safe evaluation of potential risks and preserves responsible communication.

3.2 Sleeper Code Model and Payload Design

Sleeper code refers to structured non-executable strings that appear inside protein data files. These strings resemble metadata blocks or command-like patterns. They remain inert yet may persist across workflows. The study uses three common embedding surfaces in protein systems. These include FASTA headers, CIF remark lines and interior sequence positions. The markers remain syntactically neutral. They do not resemble any known system instruction. They are easy for humans to see. They are also easy for the simulation to detect. This enables controlled measurement of propagation effects.

3.3 Synthetic Input Construction

The simulation generates synthetic FASTA and mmCIF files. FASTA files contain canonical amino acids. CIF files follow standard mmCIF formatting. Marker placement is controlled to allow measurement of persistence. These files reflect the structure of real formats. They create realistic conditions without risk. Each file contains only synthetic content. No biological meaning is implied. No sensitive sequence appears in the study. The files demonstrate structural behavior under transformation. This supports transparent evaluation of pattern movement.

3.4 Workflow Simulation Architecture

The workflow comprises five conceptual stages, as seen in modern protein science. The ingestion stage reads and normalizes protein data. The database stage forms logs or identifiers. The alignment stage prepares data for analysis. The HPC stage builds job scripts. The visualization or federation stage aggregates or displays information. These stages appear in **Table 1**. The table summarizes their roles and representative transformations. The simulation does not treat markers as instructions. It records changes in their visibility. This produces a clear trace of file evolution.

Table 1: Conceptual Model of the Simulated Workflow

Workflow Stage	Primary Purpose	Representative Transformations
Ingestion	Read and normalize protein data	Header and metadata extraction
Database Interaction	Prepare information for indexing	Log creation and identifier generation
Alignment Services	Reformat data for analysis	Wrapper creation and annotation
HPC Pipeline	Insert data into job scripts	Script templating and metadata growth
Visualization or Federation	Present or aggregate content	Display formatting and logging

The workflow does not interpret markers as instructions. It records only their movement or distortion.

3.5 AI Model Ingestion and Approximate Modeling

AI-assisted protein modeling introduces distinctive propagation effects. Many transformer models preserve the input prefix. Many also accept a variety of token patterns when token limits are respected. The study uses a safe ingestion stub to model this behavior. The stub detects markers and records amino acid anomalies. It does not generate biological content. This approach captures the key features of AI input handling. It models prefix preservation. It also models tolerance of anomalous tokens. It avoids any unsafe operation. This supports the analysis of AI side propagation.

3.6 Evaluation Metrics

We evaluate marker propagation using simple, stage-level measures derived from the simulation trace. All measures are defined over the abstract workflow stages described in Table 1 rather than tool-specific implementations. Propagation is characterised in three ways. First, the number of abstract workflow stages retaining the marker after transformation is recorded. Second, the span of propagation is measured as the number of stages between the earliest and latest detection of the marker, inclusive, regardless of whether preservation is continuous at every intermediate stage. Third, any alteration of the marker's structure during

processing, such as truncation or extension, is recorded. These measures support controlled comparison across experiments and provide a basis for the analysis that follows.

3.7 Implementation Summary

The simulation uses Python to generate files, apply transformations and detect markers. A unified Artifact class stores file content, workflow stage and marker presence. Each stage produces a new Artifact. This creates a complete trace of file evolution. The design is transparent and reproducible. All inputs are synthetic. No system level operations occur. The environment remains fully controlled. This ensures safety and clarity. It also supports open analysis. Each experiment follows a deterministic stage-by-stage transformation sequence, with marker detection and propagation metrics computed from the resulting artifact trace.

3.8 Ethical and Safety Considerations

The study follows cyberbiosecurity guidance. All markers (**Table 2**) are inert placeholders. The framework does not test live vulnerabilities. It does not attempt exploitation. It focuses on safe characterization of data movement. This supports community awareness without creating risk.

Table 2: Ten Experiments Evaluated in the Study

Test #	Description of Condition
1	Marker placed in a FASTA header
2	Marker placed in a CIF remark block
3	Marker inserted into a protein sequence
4	Marker carried into a database logging context
5	Marker exposed to workflow execution conditions without activation (Negative control)
6	Marker passed through alignment service formatting
7	Marker embedded inside a high performance computing script
8	Marker present during visualization or rendering
9	Marker used as part of AI model input during ingestion
10	Marker evaluated in federated data propagation conditions

These experiments represent common points at which structured patterns may enter or travel through scientific workflows. Each experiment uses the same simulation pipeline and support observation of how markers survive or fail at each stage. They frame the results and analysis that follows.

4. Results

4.1 Success Criteria

Each experiment is evaluated by the persistence of sleeper code markers across workflow stages. A test is successful when a marker survives at least one downstream stage without being removed or altered. A test fails when a stage blocks the marker or removes it. This criterion supports a unified view of propagation behavior. It provides a simple measure that connects directly to the design of the simulated workflow.

4.2 Experimental Outcomes

Table 3 presents the outcomes for all ten experiments, including **Test 5**, which serves as a negative control confirming that passive marker persistence does not imply execution or active compromise. The table reflects direct observations from the simulation trace for each experiment, with marker presence or absence recorded at each stage. The final outcome for each experiment appears in the Success column, allowing a clear comparison across conditions. **Figure 2** reports the stage-level propagation measures underlying **Table 3**.

Table 3: Summary of Experimental Outcomes

Test #	Category	Outcome	Rationale
1	FASTA header injection	Success	Markers persisted through all stages
2	CIF remark injection	Success	Metadata fields preserved markers fully
3	Sequence embedding	Partial Success	Only one sequence retained its marker
4	Database query propagation	Success	Query processing passed markers forward
5	Pipeline execution	Negative Control	No execution or reinterpretation occurred
6	Cloud alignment propagation	Success	Alignment wrappers preserved markers
7	HPC ingestion	Partial Success	Markers survived packaging but never activated
8	Visualization parsing	Success	Logs displayed intact marker content
9	AI model ingestion	Partial Success	ProtGPT2 preserved and extended markers
10	Federated database cascade	Success	Markers persisted to the final state

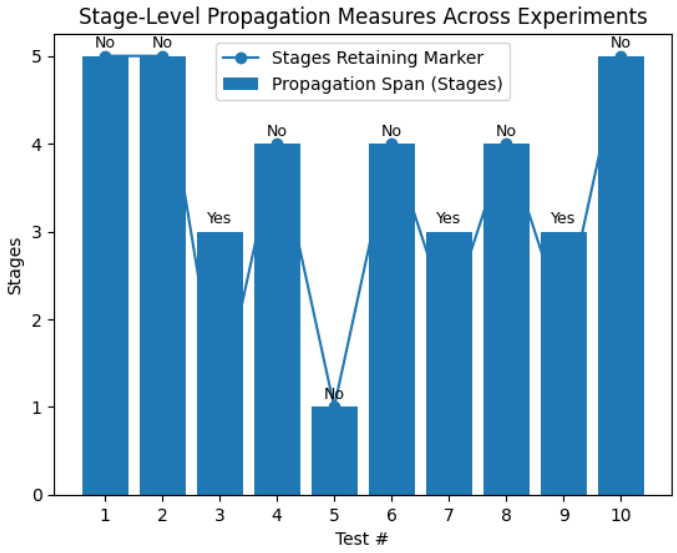


Figure 2: Stage-level propagation span and marker retention across experimental tests, with Yes/No indicating the presence or absence of observed marker distortion

These outcomes show consistent propagation through many workflow components. Several stages preserve structured patterns with no modification. Some cases produce partial results when only a subset of inputs carries markers. The single failure case shows that passive propagation does not imply execution risk. The range of outcomes across tests provides insight into the structural behavior of scientific workflows.

4.3 Interpretation of Propagation Patterns

Propagation patterns reflect each workflow stage's tolerance to unexpected text. Stages that wrap content or log content often pass markers unchanged. Stages that create new representations sometimes preserve markers from the inherited text. No stage interprets a marker as a command or instruction. This confirms the simulation's passive nature and highlights its structural features.

The results reveal that many components of protein science workflows handle data in a permissive manner. Such permissive handling allows embedded patterns to survive multiple transformations. The behavior supports the hypothesis stated in the Introduction. The findings show that structural persistence is common in these workflows. These observations create a foundation for the risk analysis that follows.

4.4 Interpretation for Test Nine

The ninth experiment examines marker propagation through AI model ingestion. The simulation reflects the known behavior of ProtGPT2 when the input remains under token limits. The model accepts a variety of tokens without validation. It also preserves prefix content and extends sequences with new material. This behavior supports a strong propagation pathway for sleeper code patterns.

These observations align with empirical findings described in the Methods. Other models may enforce tighter controls or do not require user input in the forms we have tested. Some models may also reject unexpected tokens. For this reason, the result remains partial across model families. The propagation observed in ProtGPT2 remains strong due to its permissive design.

5. Discussion

The experiments show that protein data files can function as under protected digital objects. They behave as semi-structured containers that carry symbolic content beyond biological meaning. The workflow stages allow these patterns to move across cloud services and HPC systems. They also pass through visualization tools without loss. This confirms that scientific workflows often preserve content that appears harmless.

The results highlight that many workflows rely on trust shaped by biological thinking rather than computational safety. Header fields, metadata blocks, and sequence positions accept varied patterns without scrutiny. These fields appear stable under transformation. They do not enforce constraints that prevent the inclusion of structured non-biological content. This leaves a wide surface for passive propagation.

ProtGPT2's behavior reveals a second risk. The model accepts unexpected tokens when they fall within token length limits. The model also preserves prefix content. This creates an amplification pathway for any pattern that enters the model. The pattern then persists within extended output.

The experiments do not show execution of content because the design prevents active use of markers. The absence of execution does not reduce the significance of persistence. Propagation across many stages confirms that workflows handle protein data with great permissiveness. This supports the study's central hypothesis. It also reveals a need for stronger attention to data integrity.

These findings indicate that scientific workflows may serve as silent channels for structured content. The workflows do not detect or remove these patterns during typical operations. The lack of sanitation exposes a structural weakness. This weakness may invite risk in settings where data trust is essential. It also motivates the development of new defenses for computational biology.

6. Solution Architecture

We propose a multilayer mitigation framework to address the risks identified in this study. Instead of assuming that protein data are harmless scientific artifacts, the framework treats them as potential vectors and applies cybersecurity principles suited to biological computation. Its layered controls address the silent movement of structured patterns across tools and services, creating a foundation for safer digital biology.

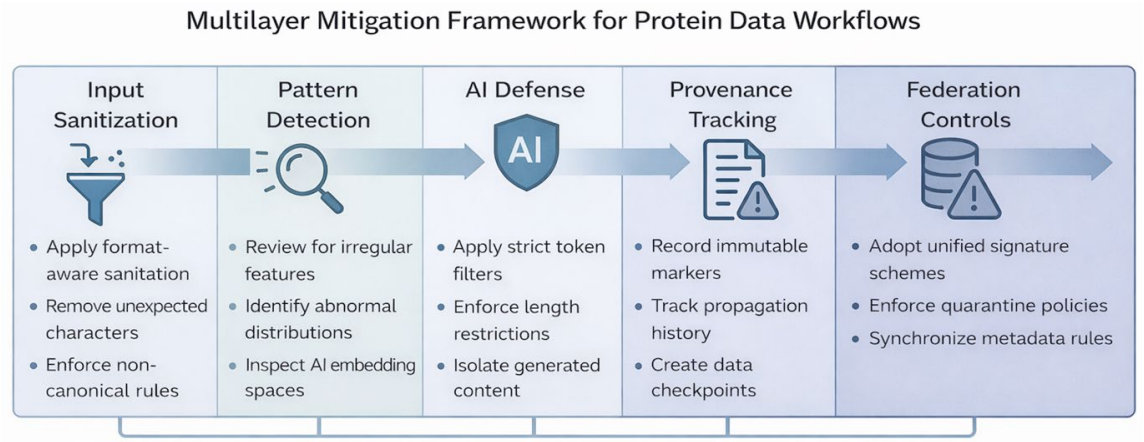


Figure 3: Multilayer mitigation framework for protein data workflows. The figure outlines five security layers that prevent, detect, and contain structured patterns as protein data move through scientific and AI enabled systems

The design adapts well known defense techniques from other security domains. Input controls, anomaly detection, provenance tracking, and containment policies serve as the core components. Although these ideas are familiar within networked environments, they have not been applied in a coordinated way to protein data workflows. Their integration here reflects the distinctive demands of computational biology.

Several differences separate this model from traditional cybersecurity approaches. It targets the quiet persistence of structured patterns within scientific formats rather than overt attempts to exploit software. AI behavior is treated as part of the attack surface. Data-level integrity becomes the focus, recognizing that scientific communities depend on shared, trusted formats. This shift places file content at the center of cyberbiosecurity practice.

The input sanitization layer forms the first line of defense. Format aware rules screen incoming files, eliminate unexpected characters in headers or metadata, and enforce biological sequence constraints when needed. Early filtering prevents unsafe text from entering downstream workflows.

Next, the anomaly detection layer evaluates structural and statistical features within files. Irregular tokens, unusual amino acid distributions, and dense metadata fields serve as signals of interest. These checks reveal content that deviates from natural protein patterns, enabling early identification of anomalous data.

The AI ingestion layer adds protections specific to machine learning systems. Token filters restrict non-standard input, length limits reduce the potential influence of complex prefixes, and separation between input and generated content prevents unwanted amplification. Prefix patterns that appear unusual can also be logged for review.

The workflow transparency layer increases visibility across the pipeline. Immutable provenance markers document each transformation. Files that contain structured patterns can be flagged, and checkpoints allow reviewers to trace how content moves through the workflow. Automated alerts help identify suspect material before it spreads.

Finally, the federation containment layer prevents anomalous content from moving across distributed ecosystems. Shared anomaly signatures and coordinated quarantine policies create consistent safeguards across repositories. Synchronized metadata rules further reduce accidental propagation and improve resilience within collaborative research networks.

Together, these layers establish defense-in-depth for protein data workflows. The framework advances awareness of data integrity risks, addresses the silent propagation observed in our experiments, and reflects the growing influence of AI in protein science. It offers a practical path for improving cyberbiosecurity across the digital biological landscape.

7. Conclusion and Future Work

Protein data files function as active digital objects rather than passive scientific records. The experiments reveal persistent movement of structured patterns across many workflow stages. Such behavior signals a clear cyberbiosecurity concern grounded in data integrity. Silent propagation across trusted systems indicates that biological formats warrant stronger oversight. These results affirm the need for new thinking in secure scientific computation.

Workflows in protein science often rely on biological trust rather than computational rigor. That trust allows symbolic patterns to persist across multiple contexts. AI models amplify this risk by permissively accepting unexpected tokens. Each of the findings exposes a structural weakness within modern bioinformatics ecosystems. Awareness of this weakness should guide future safeguards.

A multilayer mitigation framework offers a practical path toward better protection. Input controls and anomaly detection can prevent unexamined content from moving freely. Provenance tools add visibility to complex workflows. Containment measures support safety across federated systems. These measures create a workable foundation for improved resilience.

Safe simulation plays a central role in this effort. It reveals important behaviors without releasing sensitive details. It supports open inquiry while reducing potential harm. It also creates space for future studies on data-level vulnerabilities. Stronger defenses at the digital biological boundary now appear both possible and necessary.

Ethics Statement: This research was reviewed by the North Carolina Agricultural and Technical State University Institutional Review Board (IRB) with determination that this research or related activity have satisfied any relevant requirements of the university's research oversight committees to ensure ethical conduct and compliance with regulations. Compliance Committee approving the research: IRB HS23-0120 NHSR. Despite this classification, the author exercised discretion in the design and reporting of this work. Experimental methods and results that could pose biosafety or cyberbiosecurity risk were intentionally abstracted or simulated to prevent misuse. This approach reflects a commitment to responsible communication, open scientific inquiry, and the protection of scientific and public safety interests.

AI Declaration: Artificial intelligence tools were used in a limited and controlled manner in the conduct of this research. AI protein models, including ProtGPT2, and AI-enabled cloud workflows were used as research subjects to evaluate behaviors within protein data pipelines. These models were not used to generate biological discoveries, scientific conclusions, or manuscript content.

References

Alkhouri, N., Basu, S. and Kellogg, E.H. (2020) 'Exploring the predictive capabilities of AlphaFold using adversarial protein sequences', *IEEE Transactions on Artificial Intelligence*, 00(0), pp. 1–12. Available at: https://par.nsf.gov/servlets/purl/10488458

American Association for the Advancement of Science (2024) *Governance needed to ensure biosecurity of biological AI models*. AAAS News Release, August. Available at: https://www.eurekalert.org/news-releases/1054902

Carbone, A., Belingheri, M. and Bianchini, M. (2022) 'Adversarial attacks on protein language models', in *Proceedings of the Machine Learning for Structural Biology Workshop at NeurIPS 2022*. Available at: https://doi.org/10.1101/2022.10.24.513465

Department of Defense (2025) *Artificial intelligence cybersecurity risk management tailoring guide*. Version 2. Washington, DC. Available at: https://dodcio.defense.gov/Portals/0/Documents/Library/AI-CybersecurityRMTailoringGuide.pdf

Goodfellow, I.J., Shlens, J. and Szegedy, C. (2015) 'Explaining and harnessing adversarial examples', *International Conference on Learning Representations (ICLR)*. Available at: https://arxiv.org/abs/1412.6572

Gore, S. *et al.* (2017) 'Validation of structures in the Protein Data Bank', *Structure*, 25(12), pp. 1916–1927. Available at: https://doi.org/10.1016/j.str.2017.10.009

Murch, R.S., So, W., Buchholz, W.G., Raman, S. and Peccoud, J. (2020) 'Facing the 2020 pandemic: what cyberbiosecurity wants us to know to safeguard the future', *Journal of the American Biological Safety Association*, 25(2), pp. 55–67. Available at: https://doi.org/10.1016/j.bsheal.2020.09.007

National Cybersecurity Center of Excellence (2023) *Cybersecurity and privacy of genomic data*. Available at: https://www.nccoe.nist.gov/projects/cybersecurity-and-privacy-genomic-data

National Institute of Standards and Technology (2022) *Cybersecurity of genomic data*. NIST IR 8432. Available at: https://doi.org/10.6028/NIST.IR.8432

National Science Advisory Board for Biosecurity (2023) *Proposed biosecurity oversight framework for the future of science*. Available at: https://osp.od.nih.gov/wp-content/uploads/2023/03/NSABB-Final-Report-Proposed-Biosecurity-Oversight-Framework-for-the-Future-of-Science.pdf

Ney, P., Ceze, L. and Kohno, T. (2017) 'Computer security, privacy, and DNA sequencing', in *Proceedings of the 26th USENIX Security Symposium*, pp. 765–779. Available at: https://www.usenix.org/system/files/conference/usenixsecurity17/sec17-ney.pdf

Peccoud, J., Gallegos, J.E., Murch, R.S., Buchholz, W.G. and Raman, S. (2018) 'Cyberbiosecurity from naive trust to risk awareness', *Trends in Biotechnology*, 36(1), pp. 4–7. Available at: https://doi.org/10.1016/j.tibtech.2017.10.012

Pope, T. and Patooghy, A. (2025a) 'Probing AlphaFold's input attack surface via red teaming', in *Proceedings of the 22nd Annual International Conference on Privacy, Security, and Trust*, pp. 1–10. https://doi.org/10.1109/PST65910.2025.11268825

Pope, T. and Patooghy, A. (2025b) 'Structured perturbations in protein design models', *NeurIPS BioSafe GenAI Workshop*. Available at: https://openreview.net/forum?id=BItfXx4ozE

Urbina, F., Lentzos, F., Invernizzi, C. and Ekins, S. (2022) 'Dual use of artificial intelligence powered drug discovery', *Nature Machine Intelligence*, 4, pp. 189–191. Available at: https://doi.org/10.1038/s42256-022-00465-9

White House (2022) *Executive order on advancing biotechnology and biomanufacturing innovation for a sustainable safe and secure American bioeconomy*. Available at: https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2022/09/12/executive-order-on-advancing-biotechnology-and-biomanufacturing-innovation-for-a-sustainable-safe-and-secure-american-bioeconomy/

White House (2025) *Improving the safety and security of biological research*. Available at: https://www.whitehouse.gov/presidential-actions/2025/05/improving-the-safety-and-security-of-biological-research/

wwPDB Consortium (2019) 'Protein Data Bank: the single global archive for 3D macromolecular structure data', *Nucleic Acids Research*, 47(D1), pp. D520–D528. https://doi.org/10.1093/nar/gky949

Masters Research Papers

Applicability Of Industry 4.0 Technologies in Combating Fraud in the Financial System

Jennyfer da Conceição Fonseca Santos, Luciana Paula Reis and June Marques Fernandes
Universidade Federal de Ouro Preto, João Monlevade, Brasil

jennyfer.santos@aluno.ufop.edu.br
lucianapaula@ufop.edu.br
june@ufop.edu.br

Abstract: The increasing digitalization of financial services has brought significant benefits, such as speed and convenience, but it has also increased exposure to sophisticated fraud schemes, including identity theft, card fraud, and money laundering. In this context, Industry 4.0 (I4.0) technologies—particularly Machine Learning (ML) and blockchain—have emerged as strategic tools for enhancing security and mitigating risks in the banking sector. This study aims to examine the applicability of these technologies in detecting and mitigating financial fraud, addressing the following research question: How have blockchain and ML technologies contributed to mitigating different categories of fraud? To answer this question, a Systematic Literature Review (SLR) was conducted using the Scopus and Web of Science databases, focusing on open-access articles published between 2021 and 2025 that were aligned with the research topic. Following the PRISMA protocol, 27 articles were selected and analyzed through bibliometric analysis, content analysis, and mapping of future research directions. The originality of this research lies in the integration of cutting-edge ML and blockchain analyses to mitigate financial fraud, bridging technology, challenges, and gaps. Despite these advances, challenges remain, including data imbalance, the need for model interpretability, and ethical and regulatory concerns. Future research should focus on integrating ML and blockchain, improving algorithms to handle imbalanced datasets, and developing explainable and secure solutions. This study contributes by providing an up-to-date and comprehensive overview of current trends and research gaps in the application of I4.0 technologies to financial security, serving as a foundation for scientific progress and the adoption of these tools by banking institutions.

Keywords: Industry 4.0, Machine learning, Blockchain, Financial system, Fraud prevention

1. Introduction

The increasing digitalization of financial services has brought a series of benefits to both banking institutions and their customers, making transactions faster, more convenient, and more accessible (Chang *et al*., 2022). However, this transformation has also increased the banking sector's vulnerability to increasingly complex frauds, such as identity theft, credit card fraud, advanced money laundering schemes, and data manipulation (Asmar and Tuqan, 2024). According to Btoush *et al*. (2023), cybercrimes related to these transactions pose a constant challenge for financial institutions, which need to invest in more advanced techniques to detect and prevent such crimes.

Industry 4.0 (I4.0) technological innovations emerge as strategic alternatives to strengthen the security and robustness of financial systems against fraud (Ismaeil, 2024). Machine learning (ML), a subset of AI, has been widely used to detect suspicious patterns in transactions, allowing for more accurate and faster identification of fraudulent activities (Asmar and Tuqan, 2024). Approaches such as supervised and unsupervised learning are crucial for improving anomaly detection systems and reducing financial losses (Hilal et al., 2022). Furthermore, recent research indicates that combining graph databases with ML models enhances the identification of unusual behavior, increasing the effectiveness of anti-fraud systems (Patil et al., 2024). By processing large volumes of data in real time, artificial intelligence reduces false positives and improves transaction security, benefiting both customers and financial institutions (Ismaeil, 2024).

The literature identifies ML and blockchain as central to combating financial fraud, including fraudulent banking transactions, identity fraud, credit card fraud, money laundering, and document forgery (Baabdullah *et al*., 2024; El-Chaarani and El-Abiad, 2024; Patil *et al*., 2024). ML stands out for its ability to detect anomalous patterns in real time, while blockchain offers traceability, immutability, and transparency. The integration of these technologies improves detection accuracy, reduces false positives, and enables automated mitigation through smart contracts and collaborative systems, strengthening incident prevention and response in the financial sector (Rabbani *et al*., 2024; Ren *et al*., 2023). Mapping trends, limitations, and application potential contributes to the advancement of academic knowledge at the intersection of I4.0 and fraud prevention.

This article aims to analyze the applicability of I4.0 technologies, with a focus on ML and blockchain, in treating and mitigating the effects of fraud in the banking system. The study begins with the following question: "What is the current progress of academic research related to the application of these technologies in financial fraud

mitigation?" To achieve this objective, this study adopts a Systematic Literature Review (SLR), gathering data through a bibliographic search. The bibliographic search aims to collect, identify, select, and critically evaluate previously published references in various forms of academic and scientific materials (Marconi and Lakatos, 2021).

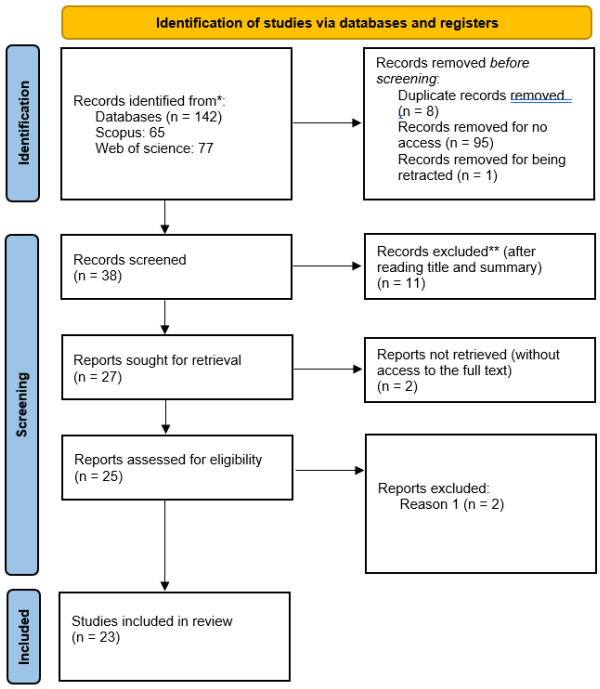
Existing literature predominantly focuses on the use of blockchain or ML for fraud mitigation, without, however, comprehensively exploring the complementarity and synergy between these and other I4.0 technologies. Therefore, this SLR aims to fill this gap by providing a deeper and more up-to-date understanding of the state of research and implementation practices of these technologies in combating banking fraud.

This research is divided into four sections in addition to this brief introduction. Section 2 presents the methodology used to execute and support this work. Section 3 presents the main results and discussions of the analyses performed. Finally, Section 4 presents the final considerations obtained from the research, the limitations of the work, and proposals for future studies.

2. Methodology

The method employed here is theoretical-conceptual, utilizing the SLR tool. SLR is a rigorous methodology that synthesizes scientific evidence from multiple studies, providing a comprehensive and reliable overview of a specific topic (Roever, 2020). To conduct the SLR, the PRISMA principles (Page *et al*., 2021) were used to search for, identify, and select articles for inclusion in the research.

The Scopus and Web of Science databases were selected as data sources, as they are two of the largest and most comprehensive academic databases, containing a wide range of journals from various fields. The search was conducted in the "Title, Abstract, and Keywords" fields, using the keywords "Fraud," "Technologies," "Security," and "Finance." Restrictions were applied to include only articles published between 2021 and 2025, aiming to understand the current discussions on the topic over the past five years. Furthermore, only open access documents were selected for analysis. Figure 1 presents the Prisma flowchart used in the research:



Source: Adapted from Page *et al*. (2021).

Figure 1: PRISMA flowchart for article identification and screening

After conducting the database search and applying the filtering criteria, the search yielded a total of 142 articles. After searching the database, 8 duplicate articles, 95 articles that were not open access, and 1 article that had been retracted by the journal were removed, resulting in 38 selected articles. Subsequently, the titles and abstracts were read to exclude articles that did not fit the research topic, resulting in the exclusion of 11 articles. Subsequently, after reading the full article, 2 articles whose full texts were not found were excluded, as well as 2 articles that were also not aligned with the research topic. The final result of this review was 23 articles.

Three distinct analyses were conducted to understand and synthesize the existing body of knowledge on the topic in question: i) bibliometric analysis, ii) content analysis, and iii) analysis of suggested future work. The bibliometric analysis included identifying and categorizing articles according to various criteria, such as year of publication and first author's country, as well as creating a keyword cloud using VOSviewer software. Content analysis focused on the themes and subthemes discussed in the articles. Finally, analyzing the recommendations for future work described in the articles provides valuable insights into future research directions in this specific field. These combined analyses offer a comprehensive and in-depth understanding of the current state and emerging trends in technologies used to mitigate fraud in financial systems.

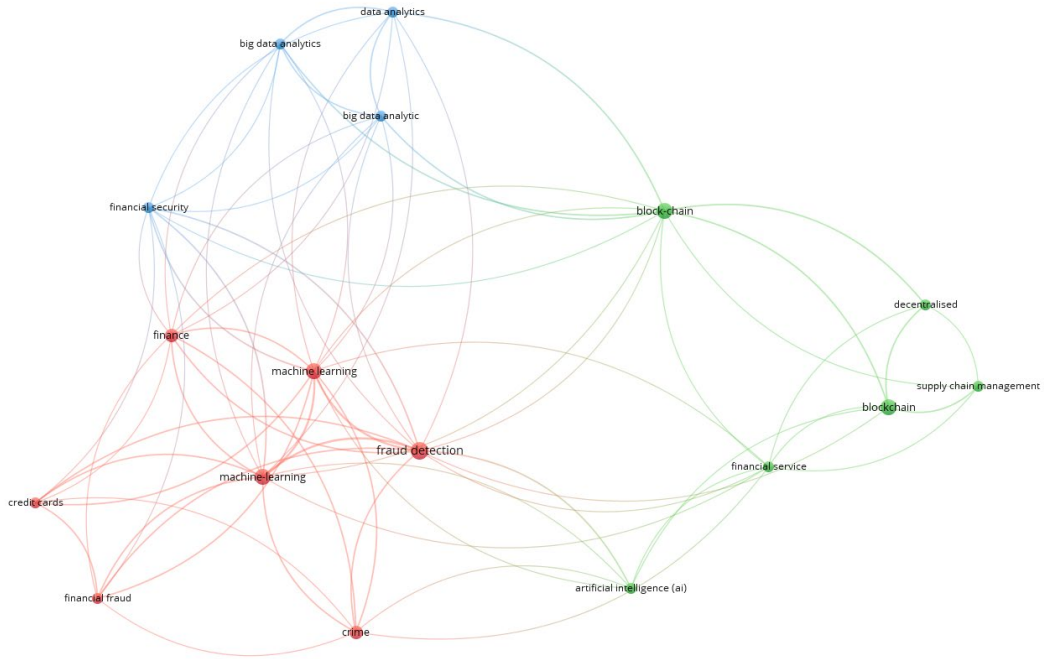
3. Results and Analysis

This section presents the results of the SLR analyses, containing three subsections: bibliometric analysis; content analysis; and analysis of future work.

3.1 Bibliometric Analysis

The analysis of the number of articles published per year reveals an increasing trend in academic production over time. In 2021 and 2023, there was an equal number of publications (4 articles each), while in 2022 there was a significant increase to 6 articles. However, the most significant growth occurred in 2024, with 13 publications, representing a peak in production.

The progressive increase in publications may indicate a growing academic interest in the topic, as well as a greater availability of related research and studies. The increase in publications in 2024 suggests that the topic may be gaining relevance in the scientific community, possibly driven by technological advances or greater awareness of its importance. If this trend continues, it is possible that an even greater volume of articles will be published in the coming years, consolidating the area as a growing field of study. An analysis of articles by the first author's country of origin reveals a heterogeneous distribution across different regions, as shown in the figure. India stands out as the country with the highest number of publications (9), indicating strong academic production in the field. China follows with 3 publications, demonstrating a significant but significantly lower share compared to India. Other countries, such as Saudi Arabia, Romania, the United Arab Emirates, the United States, Malaysia, Nigeria, and the United Kingdom, have a moderate number of publications (2 each), suggesting a significant but less significant share. On the other hand, Brazil appears with only one publication, representing a smaller contribution to the sample analyzed. This distribution may reflect different levels of research investment, academic interest in the topic, or access to resources and incentives for scientific production and publication.



Source: Data tabulated in VOSviewer.

Figure 2: Word Cloud

Word cloud analysis identifies three major thematic clusters that stand out in the literature. The first cluster is directly related to fraud detection in the financial sector, with terms such as fraud detection, ML, crime, financial fraud, credit cards, and finance. This cluster demonstrates that the application of artificial intelligence techniques, especially ML, is central to strategies aimed at combating financial fraud, often associated with transaction security and crime prevention. The second cluster has as its central element the term blockchain, which connects to topics such as financial service, decentralized, supply chain management, and artificial intelligence (AI). This cluster highlights the growing use of blockchain technology not only in the financial sector but also in solutions aimed at decentralizing processes and managing supply chains, reinforcing its role in promoting data security, transparency, and traceability. The third cluster focuses on the terms big data analytics and financial security. This group reflects the importance of large-scale data analysis tools as essential support for financial security, enabling the identification of patterns, the prediction of suspicious behavior, and the generation of insights for risk mitigation.

In general, the topics are strongly interconnected, particularly through the terms fraud detection, ML, blockchain, and financial security. This demonstrates that current research adopts a multidisciplinary approach, integrating emerging technologies to address the challenges related to financial security, fraud detection, and innovation in financial services.

3.2 Content Analysis

The selected studies cover a wide range of application areas. They investigate issues such as cybersecurity in the financial sector, fraud detection in banking transactions, and the use of emerging technologies to improve financial crime monitoring and prevention systems.

3.2.1 Fraud in the financial system

Fraud is defined as an intentional act of deception aimed at obtaining an undue advantage or causing harm to third parties. It is characterized as a non-violent crime, but highly damaging in the economic and corporate context (Kulmie *et al*., 2024). The literature identifies two main groups of frauds: internal and external.

Internal fraud in the banking sector is a complex and multifaceted problem, committed by employees or managers within the organization, often exploiting flaws in internal control and corporate governance systems (El-Chaarani and El-Abiad, 2022; Kyrychenko *et al*., 2021; Bonsu *et al*., 2018). These frauds can be especially damaging because they involve the use of insider privileges to divert resources or compromise the organization's integrity. They occur due to factors such as a lack of effective anti-fraud policies, inadequate training, insufficient compensation, and inexperienced personnel (Kulmie *et al*., 2024; Bonsu *et al*., 2018).

External frauds are committed by individuals or groups outside the financial institution and can include a wide variety of attacks, many of which exploit technological vulnerabilities or psychologically manipulate victims (Taher *et al*., 2024; Btoush *et al*., 2023). Attacks range from direct data theft to more complex methods such as phishing and social engineering, which involve psychologically manipulating users to obtain confidential information (Taher *et al*., 2024). Cyberfraud with credit cards is one of the most prevalent forms, costing banking institutions billions of dollars annually and involving the unauthorized use of a card or its information to carry out fraudulent transactions (Baabdullah *et al*., 2024; Charizanos *et al*., 2024). Furthermore, practices such as money laundering and identity theft pose ongoing risks that exploit flaws in internal control systems and corporate governance (Kulmie and Ibrahim, 2024). Regarding the most frequently discussed types of fraud, general banking transactions lead, being addressed in sixteen papers (Baabdullah *et al*., 2024; Ghrib *et al*., 2024; Patil *et al*., 2024; Rabbani *et al*., 2024; Wei and Lee, 2024; Chang *et al*., 2022; Aljofey *et al*., 2022; Kyrychenko *et al*., 2021; Ileberi *et al*., 2021). Next, identity fraud was discussed in 10 articles (Ali *et al*., 2024; Asmar and Tugan, 2024; Ismaeil, 2024), reflecting growing concern about the risks associated with the theft and misuse of personal data. Credit card fraud and money laundering are also among the most discussed topics, each appearing in 9 studies (El-Chaarani and El-Abiad, 2024; Patil *et al*., 2024; Wijaya *et al*., 2024). With the advancement of I4.0, financial fraud detection requires new technological paradigms to respond to growing risks. I4.0 technologies emerge as strategic solutions to strengthen the security and resilience of financial systems against fraud (Ismaeil, 2024).

3.2.2 Technologies to combat fraud

With the advancement of I4.0, financial fraud detection requires new technological paradigms to respond to growing risks. Technologies such as AI, ML, big data, and blockchain have shown promise for detecting these frauds and preventing cyberattacks (Patil *et al*., 2024). The articles analyzed for this systematic literature review

reveal a clear predominance of ML use, present in 21 articles. Blockchain also stands out, present in 20 studies, frequently combined with ML, Federated Learning (FL) (ML technique that collaboratively trains models across different institutions or devices, without sharing the raw data, ensuring privacy and security), or data mining techniques, especially in cases requiring transparency and immutability, such as money laundering and distributed denial-of-service (DDoS) attacks (Ahmed and Alabi, 2024; Baabdullah *et al*., 2024; Wei and Lee, 2024; Aljofey *et al*., 2022; Chang *et al*., 2022). On the other hand, emerging technologies or those with more specific uses, such as FL, graph technologies, deep learning, biometrics, and data balancing techniques, appear less frequently, indicating a still incipient adoption trend or potential implementation challenges.

Detection-focused approaches predominate over mitigation strategies. The literature has shown greater interest in the application of technologies capable of identifying anomalous patterns, suspicious behavior, or early signs of irregularities (Ahmed and Alabi, 2024; Gaikwad *et al*., 2023; Ren *et al*., 2023; Aljofey *et al*., 2022; Ashfaq *et al*., 2022; Chang *et al*., 2022). In contrast, mitigation strategies appear less systematically (Ali *et al*., 2024; Asmar and Tugan, 2024; Kyrychenko *et al*., 2021), suggesting a gap in the structuring of mechanisms capable of containing the impacts of fraud or adapting organizational processes in the face of already materialized threats (El-Chaarani and El-Abiad, 2024; Khan *et al*., 2023). Table 1 presents the applications for ML and blockchain, technologies that appear most frequently in the literature.

Table 1: Applications of ML and blockchain

Technology	Applications	Authors
Machine Learning (ML)		
Supervised Learning	It uses algorithms such as Logistic Regression, Random Forest, and Convolutional Neural Networks (CNN) to detect card fraud, classify transactions, and correct data imbalances. It is also used to identify accounting fraud and suspicious behavior in real time.	Ileberi et al. (2024); Ismaeil (2024); Wijaya et al. (2024), Lokanan e Sharma (2022), Stojanovic e Bozic (2022)
Unsupervised Learning	It uses algorithms such as SVM, HMM, and Autoencoders to classify transactions, detect temporal patterns, model sequences of fraudulent actions, and generate synthetic data for training.	Asmar e Tuqan (2024), Hilal et al. (2022)
Hybrid ML Techniques	They combine multiple algorithms (e.g., CNN, RNN, LSTM) to detect complex and sequential patterns, especially in time series, and map connections between users and transactions.	Patil et al. (2024), Wei e Lee (2024), Ren et al. (2023)
Federated Learning	It allows multiple participants to collaborate on model training without sharing their raw data, ensuring privacy. It is combined with techniques such as SVM and neural networks.	Rabbani et al. (2024), Ahmed e Alabi (2024)
Blockchain		
Blockchain (isolated)	Ensures the traceability and integrity of financial records through its immutable nature, which helps reduce corruption, money laundering, and human error.	Kulmig et al. (2024), Abad-Segura et al. (2024)
Blockchain + ML / Deep Learning	The combination enhances high-precision fraud detection by analyzing patterns in large volumes of transaction data, including temporal anomalies.	Aldabam e Hamdi (2024), Gałkoyad et al. (2023)
Blockchain + Federated Learning	Creates collaborative models for real-time fraud detection while maintaining the privacy of participating financial institutions' data.	Rabbani et al. (2024), Baabdullah et al. (2024)
Blockchain + Other Technologies	When combined with GBDT, it improves accuracy on imbalanced data. With the Zero Trust model, it offers continuous authentication. Combined with mobile protection, it combats phishing and vishing.	Paşek e Sharma (2024), Njoya et al. (2023), Ren et al. (2023)

Source: The authors.

ML focuses on the development of algorithms and computational models capable of learning patterns from data, without being explicitly programmed for a specific task (Asmar and Tuqan, 2024; Ismaeil, 2024). It has been widely applied in cybersecurity, financial fraud detection, banking transaction analysis, and data protection, demonstrating the ability to handle large volumes of information and identify complex patterns that are difficult to detect manually (Asmar and Tuqan, 2024; Wei and Lee, 2024; Gaikwad *et al*., 2023). These systems use historical data to "learn" and improve their performance over time, becoming increasingly accurate in tasks such

as classification, prediction, and anomaly detection (Ileberi and Sun, 2024; Wijaya *et al*., 2024). Algorithms can be supervised (trained with labeled data), unsupervised (identify patterns in unlabeled data), or semi-supervised (combine both approaches) (Ismaeil, 2024; Hilal *et al*., 2022). Furthermore, the integration of ML techniques with technologies such as blockchain and federated learning has further expanded their applications, ensuring greater security, privacy, and efficiency in financial and technological systems (Rabbani *et al*., 2024; Ren *et al*., 2023).

The use of blockchain to monitor and record financial transactions offers a transparent and secure platform, enabling collaboration between financial institutions without the need to exchange sensitive data. This offers advantages related to transaction traceability and reduced corruption and money laundering (Rabbani *et al*., 2024). However, when used in isolation, it fails to detect complex fraud patterns and is limited to post-facto auditing (Kulmie *et al*., 2024). The integration of blockchain with ML/deep learning significantly expands its potential, enabling highly accurate fraud detection (Gałkowad *et al*., 2023).

ML complements the immutability of blockchain by enabling the joint analysis of on-chain data (recorded on the network) and off-chain data (such as banking histories or digital interactions). While blockchain ensures the traceability and integrity of transactions, it cannot, by itself, identify hidden patterns of fraud (Ismaeil, 2024; Asmar and Tuqan, 2024; Wei and Lee, 2024; Gaikwad et al., 2023; Ren et al., 2023; Rabbani et al., 2024; Ahmed and Alabi, 2024). In this sense, ML algorithms, such as recurrent neural networks and hybrid methods, are capable of correlating multiple sources of information, detecting anomalous behaviors and sophisticated fraud schemes that exploit gaps between the internal and external environments of the blockchain (Ren et al., 2023; Gaikwad et al., 2023; Wei and Lee, 2024). This integration expands predictive and preventive capabilities, allowing complex attacks to be identified before they consolidate.

The analyzed literature highlights significant advances in the use of I4.0 technologies to combat financial fraud, with an emphasis on ML techniques (Patil *et al*., 2024; Wijaya *et al*., 2024; Gaikwad *et al*., 2023) and blockchain-based solutions (Baabdullah *et al*., 2024; Rabbani *et al*., 2024). Recent studies highlight the predominance of approaches aimed at automated fraud detection, particularly by identifying behavioral patterns and anomalies in banking transactions, using algorithms such as Random Forest, XGBoost, and convolutional neural networks (Asmar and Tuqan, 2024; Wijaya *et al*., 2024). Despite the sophistication of these approaches, there is a significant gap in strategies aimed at mitigating and responding to fraud. Most studies focus on predictive models aimed at detection, but do not delve into the practical implications of adopting these technologies within the banking environment (El-Chaarani and El-Abiad, 2024; Kyrychenko *et al*., 2021). Furthermore, although some research mentions the integration of ML with blockchain to improve security and traceability (Gaikwad *et al*., 2023; Ren *et al*., 2023), the applicability of these systems in concrete operational contexts still lacks robust empirical evidence.

Another critical point concerns the scarcity of empirical studies based on real data from financial institutions. Much of the academic literature uses public or simulated databases, limiting understanding of the operational and regulatory barriers involved in the practical application of these technologies (Lescano-Delgado, 2023; Stojanović *et al*., 2021). Furthermore, there is little articulation between different types of fraud and the specificity of the most appropriate technological solutions for each case (Kulmie *et al*., 2024; Chang *et al*., 2022).

3.3 Directions for Future Research

The reviewed studies point to several opportunities for future research in financial fraud detection, ranging from the application of new technologies to overcoming methodological, technical, and ethical challenges. One of the main directions in Stojanović and Božić's (2022) research involves expanding the risk assessment architecture in fraud detection systems, considering broader cyber threats and vulnerabilities. Furthermore, the implementation of new security properties would contribute to strengthening the reliability of anomaly detection-based systems in the financial sector.

Another relevant aspect concerns the improvement of ML algorithms, particularly in addressing data imbalance in fraudulent transaction sets, which can compromise detection effectiveness. Strategies such as data augmentation and the development of robust adversarial countermeasure models are indicated as promising avenues for strengthening predictive accuracy (Ashfaq *et al*., 2022). Furthermore, testing ML algorithms from other domains and collecting new financial datasets can improve the applicability of these solutions (Stojanović *et al*., 2021).

Methodologically, combining neural networks with decision tree-based techniques, such as Gradient Boosting Decision Trees (GBDT), has been identified as a promising alternative for dealing with highly sparse data (Ren *et*

al., 2023). Furthermore, hyperparameter tuning and careful feature selection can optimize model training time without compromising detection accuracy (Ileberi and Sun, 2024).

Beyond the technical challenges, the application of AI in financial security raises ethical and regulatory concerns. Issues such as algorithmic bias, data privacy, and regulatory compliance need to be examined to ensure that the implementation of these solutions occurs in a fair and transparent manner (Asmar and Tuqan, 2024). Financial institutions should invest in high-quality data and strategies that ensure the reliability of AI-based systems, in addition to adapting to emerging regulations (Ismaeil, 2024).

Finally, future research could focus on developing testable hypotheses about fraudster behavior in the financial market, exploring interdisciplinary approaches that integrate criminological theories with ML techniques (Lokanan and Sharma, 2022). It is also crucial to compare different ML-based cybersecurity methods to assess their effectiveness in real-world scenarios (Asmar and Tuqan, 2024).

4. Final Considerations

This study provided a detailed analysis of the current state of research on the application of technologies to financial fraud mitigation, highlighting advances, emerging trends, and future directions for the field. Through SLR, it was possible to identify a growing academic production on the topic, with a significant increase in publications starting in 2024, reflecting the growing interest and relevance of these technologies in the financial sector.

The literature highlights that the use of ML has proven highly effective in detecting anomalous patterns in financial transactions, while blockchain technology stands out as a robust solution for ensuring the transparency and security of transactions. However, despite these advances, studies have also revealed persistent challenges, such as data imbalance in transaction sets, the need for greater interpretability of AI models, and the ethical and regulatory issues involved in the application of these technologies.

Directions for future research indicate that the development of new ML techniques, improving model interpretability, and exploring the use of blockchain in permissioned networks will be fundamental to advancing the field. Furthermore, it is essential that future research integrate interdisciplinary approaches, considering the ethical, regulatory, and practical aspects of implementing these technologies in financial institutions. In short, the application of blockchain and ML to financial fraud mitigation offers great potential to transform the security of banking systems, but continued investment in research that addresses technical and ethical challenges is necessary, promoting increasingly effective and fair solutions.

The main contribution of this study is to provide a comprehensive and up-to-date overview of the use of I4.0 technologies in financial fraud mitigation, highlighting emerging trends and technological advances. The study also offers a critical reflection on the integration of these technologies in the financial sector, contributing to a deeper understanding of cybersecurity in the banking context.

This research has limitations related to the method used. These limitations stem from the search terms selected and the inclusion/exclusion criteria of the articles, which may result in bias on the part of the researchers and the possible exclusion of other relevant work. For future research, we suggest including new terms that can encompass different aspects of the research topic, such as potential benefits and implementation barriers. Furthermore, we suggest considering the applicability of these technologies in different contexts and types of financial institutions.

Acknowledgements

The authors would like to thank the Universidade Federal de Ouro Preto (UFOP/Brazil) (www.ufop.br), the Pró-Reitoria de Pesquisa, Pós-graduação e Inovação (PROPPI), the Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior (Capes), and the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), for their support during the development of this research.

Ethics Statement: This research did not involve human or animal participants, nor did it collect sensitive data that required approval from a research ethics committee. Therefore, ethics approval was not considered necessary for this study.

AI Use Statement: In the creation of this article, artificial intelligence tools were used.

AI tools, specifically Google Gemini and GPT-4, were employed to rephrase sentences for clarity, check grammar and spelling, and summarize sections of the literature. All final content and conclusions presented are the responsibility of the authors.

References

Ahmed, A. A., and Alabi, O. (2024). Secure and scalable blockchain-based federated learning for cryptocurrency fraud detection: A systematic review. *IEEE Access*.

Ali, G., Mijwil, M. M., Buruga, B. A., and Abotaleb, M. (2024). A comprehensive review on cybersecurity issues and their mitigation measures in FinTech.

Aljofey, A., Rasool, A., Jiang, Q., and Qu, Q. (2022). A feature-based robust method for abnormal contracts detection in ethereum blockchain. *Electronics*, *11*(18), 2937.

Ashfaq, T., Khalid, R., Yahaya, A. S., Aslam, S., Azar, A. T., Alsafari, S., and Hameed, I. A. (2022). A ML and blockchain based efficient fraud detection mechanism. *Sensors*, *22*(19), 7162.

Asmar, M., and Tuqan, A. (2024). Integrating ML for sustaining cybersecurity in digital banks. *Heliyon*, *10*(17).

Baabdullah, T., Alzahrani, A., Rawat, D. B., and Liu, C. (2024). Efficiency of Federated Learning and Blockchain in Preserving Privacy and Enhancing the Performance of Credit Card Fraud Detection (CCFD) Systems. *Future Internet*, *16*(6), 196.

Bonsu, O. A. M., Dui, L. K., Muyun, Z., Asare, E. K., and Amankwaa, I. A. (2018). Corporate fraud: Causes, effects, and deterrence on financial institutions in Ghana. *European Scientific Journal*, *14*(28), 315-335.

Btoush, E. A. L. M., Zhou, X., Gururajan, R., Chan, K. C., Genrich, R., and Sankaran, P. (2023). A systematic review of literature on credit card cyber fraud detection using machine and deep learning. *PeerJ Computer Science*, *9*, e1278.

Chang, V., Di Stefano, A., Sun, Z., and Fortino, G. (2022). Digital payment fraud detection methods in digital ages and Industry 4.0. *Computers and Electrical Engineering*, *100*, 107734.

Charizanos, G., Demirhan, H., & İçen, D. (2024). An online fuzzy fraud detection framework for credit card transactions. Expert Systems with Applications, 252, 124127.

El-Chaarani, H., and El-Abiad, Z. (2024). The impact of public legal protection on the internal corporate governance efficiency in banking sector. *Journal of Economic and Administrative Sciences*, *40*(3), 482-515.

Gaikwad, V., Meher, K., Dass, R., Jonista, A. S., D'Souza, J., and Victor, R. (2023). Fraud Detection Using ML and Blockchain. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(6s), 584-590.

Ghrib, T., Khaldi, Y., Pandey, P. S., and Abusal, Y. A. (2024). Advanced Fraud Detection In Card-Based Financial Systems Using A Bidirectional Lstm-Gru Ensemble Model. *Applied Computer Science*, *20*(3), 51-66.

Hilal, W., Gadsden, S. A., and Yawney, J. (2022). Financial fraud: a review of anomaly detection techniques and recent advances. *Expert systems With applications*, *193*, 116429.

Ileberi, E., and Sun, Y. (2024). Advancing Model Performance With ADASYN and Recurrent Feature Elimination and Cross-Validation in ML-Assisted Credit Card Fraud Detection: A Comparative Analysis. *IEEE Access*.

Ileberi, E., Sun, Y., and Wang, Z. (2021). Performance evaluation of ML methods for credit card fraud detection using SMOTE and AdaBoost. *Ieee Access*, *9*, 165286-165294.

Ismaeil, M. K. A. (2024). Harnessing AI for Next-Generation Financial Fraud Detection: A DataDriven Revolution. *Journal of Ecohumanism*, *3*(7), 811-821.

Khan, H. U., Malik, M. Z., Nazir, S., and Khan, F. (2023). Utilizing bio metric system for enhancing cyber security in banking sector: A systematic analysis. *Ieee Access*, *11*, 80181-80198.

Kyrychenko, V., Soldatenko, O. A., Gorokhovska, O. V., Voloshyna, M. O., and Maksymova, L. O. (2021). Fraud in the banking system of Ukraine: ways to combat taking into account foreign experience. *Amazonia Investiga*, *10*(45), 208-220.

Kulmie, D. A., and Ibrahim, M. S. (2024). Bank corporate governance: Shield against fraud. *Journal of Ecohumanism*, *3*(3), 1917-1932.

Lescano-Delgado, M. (2023). Avances en el uso de inteligencia artificial para la mejora del control y la detección de fraudes en organizaciones. *Revista Científica de Sistemas e Informática*, *3*(1), e494-e494.

Lokanan, M. E., and Sharma, K. (2022). Fraud prediction using ML: The case of investment advisors in Canada. *ML with Applications*, *8*, 100269.

Marconi, M. A., and Lakatos, E. M. (2017). Fundamentals of Scientific Methodology. 8th ed. São Paulo, SP: Atlas. 346.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... and Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *bmj*, *372*.

Patil, A., Mahajan, S., Menpara, J., Wagle, S., Pareek, P., and Kotecha, K. (2024). Enhancing fraud detection in banking by integration of graph databases with ML. *MethodsX*, *12*, 102683.

Rabbani, H., Shahid, M. F., Khanzada, T. J. S., Siddiqui, S., Jamjoom, M. M., Ashari, R. B., ... and Nooruddin, M. (2024). Enhancing security in financial transactions: a novel blockchain-based federated learning framework for detecting counterfeit data in fintech. *PeerJ Computer Science*, *10*, e2280.

Ren, Y., Ren, Y., Tian, H., Song, W., and Yang, Y. (2023). Improving transaction safety via anti-fraud protection based on blockchain. *Connection Science*, *35*(1), 2163983.

Roever, L. (2020). Practical guide to systematic reviews and meta-analysis.

Stojanović, B., and Božić, J. (2022). Robust financial fraud alerting system based in the cloud environment. *Sensors*, *22*(23), 9461.

Stojanović, B., Božić, J., Hofer-Schmitz, K., Nahrgang, K., Weber, A., Badii, A., ... and Runevic, J. (2021). Follow the trail: ML for fraud detection in Fintech applications. *Sensors*, *21*(5), 1594.

Taher, S. S., Ameen, S. Y., & Ahmed, J. A. (2024). Advanced fraud detection in blockchain transactions: An ensemble learning and explainable ai approach. *Engineering, Technology & Applied Science Research*, *14*(1), 12822-12830.

Wei, S., and Lee, S. (2024). Financial anti-fraud based on dual-channel graph attention network. *Journal of Theoretical and Applied Electronic Commerce Research*, *19*(1), 297-314.

Wijaya, M. G., Pinaringgi, M. F., and Zakiyyah, A. Y. (2024). Comparative Analysis of ML Algorithms and Data Balancing Techniques for Credit Card Fraud Detection. *Procedia Computer Science*, *245*, 677-688.

Doctrine-to-Deployment: Role of Advanced Persistent Threats in Russia's Information Confrontation Doctrine

Elia Gelati[1] and Luigi Martino[2]
[1]Center for Cyber Security and International Relations Studies, University of Florence, Italy
[2]University of Bologna, Italy

eliagelati@gmail.com
luigi.martino3@unibo.it

Abstract: In 2022, before the invasion of Ukraine, many analysts feared the possibility of Russian cyberattacks overwhelming Kyiv's command and control systems and critical infrastructures, plunging Ukraine into darkness and facilitating a ground invasion. The leading entities behind these operations are advanced persistent threat groups: experienced, well-funded cyberspace actors often enjoying State sponsorship. Russian-sponsored APTs are already known for impactful cyberattacks such as the 2015 disruption of the Ukrainian energy grid, and are employed under the direction of the Russian intelligence services. Analysed through the lens of offensive realism, cyberspace appears as a domain of persistent competition among great powers. What this school of thought also posits is that while not revolutionary, cyberattacks can be used for tactical advantages. This stands in contrast with the expectations set by the Russian doctrine of "information confrontation" (or IPb), a comprehensive approach that utilises cyberattacks to achieve political, economic, and military objectives during both peacetime and wartime, with significant investments in APT groups and operations. This research analyses this "doctrine-to-deployment" gap in the role of APT groups within IPb through an offensive realist theoretical lens. It does so by analysing three case-studies of Russian APT operations using a "structured, focused" comparison methodology. These include the 2015 attack on the Ukrainian energy grid, the 2020 data breach on the SolarWinds supply chain and the APT campaigns in the war in Ukraine. By providing a doctrine-to-deployment analysis of APT units within IPb, this research clarifies a lesser-known aspect of cyber warfare: the role of State-sponsored APTs under Russian command. The results indicate that these operations, albeit aligned with State strategic objectives and the doctrine of IPb, do fall short of strategic gains, functioning more as tactical instruments of persistent warfare across wartime and peacetime. In this, the article contributes to the literature by clarifying how APTs function within IPb, bridging Russian military doctrine with international relations theory.

Keywords: Advanced persistent threat (APT) groups, Information confrontation, Russian cyber doctrine, Ukraine conflict, Cyber warfare, Offensive realism

1. Introduction

Cyberspace emerged as the fifth domain of conflict in the 2010s, alongside traditional dimensions of land, sea, air, and space. It is a contested domain, where State and non-state actors compete for strategic advantage (Allen and Gilbert, 2009). Three structural characteristics distinguish cyberspace from conventional domains: attribution ambiguity, low-cost offensives, and weak international governance (Martino, 2018), all of which contribute to making cyberspace an arena of continuous, below-threshold competition among States (Kello, 2017).

Among these actors, Russia is one of the most aggressive and sophisticated ones (Lynch III, 2025), being responsible for some of the most impactful cyberattacks in history, such as the one on the Ukrainian power grid in 2015. Over the years, Moscow codified its objectives and principles for cyberspace within the doctrine of "information confrontation" (or IPb, from the Russian "informatsionnoye protivoborstvo"), merging cyberattacks and information operations for the pursuit of its political hegemony in what it considers to be a domain of continuous conflict (Giles, 2016; Grisé *et al.*, 2022). Between the many assets at its disposal within IPb, Moscow invests funds and resources in advanced persistent threats (or APT) groups. From an offensive realist analytical perspective (Mearsheimer, 2001), these serve as tools to maximise power and ultimately ensure the State's survival in an anarchic environment characterised by intense power competition (Katagiri, 2024).

This research asks what the role of APT groups is within information confrontation, under the lens of offensive realism.

Indeed, previous policy and academic literature has only partly addressed this question. More specifically, works such as Grisé *et al.* (2022), addressed the specific topic of contrasting IPb definitions, while authors Lilly and Cheravitch (2020) and Cheravitch (2021) focused only on specific units, with their work predating active deployment of APT groups in an active warfare scenario; finally, more recently, the question was also tackled

by Voo and Singh (2025), with their efforts lacking a specific theoretical lens and a methodology to address the case-studies and derive relevant findings on the specific role, whether strategic or tactic, of APT groups.

This research addresses these issues by delivering a doctrine-to-deployment analysis of the role of APT groups within IPb, deriving hypotheses from an offensive realist theoretical framework, and then testing them through three case-studies of impactful Russian-sponsored cyber operations conducted by APT groups. For this part, the "structured, focused" methodology will be employed (George and Bennett, 2005).

This paper proceeds in six sections. Section 2 establishes the theoretical framework of offensive realism; Section 3 outlines the methodology; Section 4 traces the evolution of the IPb doctrine and maps its APT ecosystem; Section 5 analyses three case-studies that demonstrate APT operations in practice: the 2015 Ukrainian power grid attack, the 2020 SolarWinds hack, and the broader deployment of APT units during the war in Ukraine (2022-). Finally, Section 6 analyses the findings and assesses what these operations reveal about the APT roles within the Russian IPb strategy.

2. Theoretical Framework

Offensive realism is an international relations school of thought that posits that the international world order is one where States are the primary actors, among which distrust of intentions is one of the main drivers of conflict, as State strive for hegemony in order to maximize their chance of survival in what is essentially an anarchical system, without laws or order besides the one States make themselves (Mearsheimer, 2001).

The realist school has often been applied to cyberspace as a result of its similar condition of anarchy and high competition (Eriksson and Newlove-Eriksson, 2021). However, most realist do not think of cyber warfare as a game-changer; rather, it is considered a tool in the arsenal of States to be used in pursuit of its tactical objectives (Gartzke, 2013; Rid, 2013; Smeets, 2018).

Indeed, cyberspace perfectly fits some of the underlying logic of realism through three main characteristics. Firstly, while non-State actors enjoy considerable fortune in the fifth domain on the account of the lower cost of offensives (Nye, 2011), States are still the dominant actors with access to considerably more resources, which allow them to launch more precise and impactful operations against more defended targets (Gartzke, 2013; Lindsay, 2013; Smeets, 2022). Secondly, difficulties in attributing cyberattacks leads to higher uncertainty among States' and their intentions, which in turn fosters continuous competition (Kello, 2017). Thirdly, the lack of a unified and enforceable governance system renders this domain anarchical in nature (Kello, 2022). All of these characteristics transform cyberspace into an arena of persistent competition, where great powers must seek to maximize power for the sake of their survival (Lynch III, 2025).

To do so, and to offset the balance of power, they must use any tool at their disposal, such as advanced persistent threat (APT) groups (Katagiri, 2024). These are mostly State-sponsored cyber units (Steffens, 2020) that are among the most impactful actors within cyberspace, being responsible for cyber operations such as disruption of critical infrastructures or exfiltration of critical information.

As a revisionist power, Russia frequently makes use of these APT groups (Katagiri, 2024), among other cyber forces such as patriotic hackers and institutionalised troll farms (Abrams, 2016). It does so, according to the doctrine of "information confrontation", a "whole-of-government" approach that passes through the information environment and the cyber domain to reach political, military and economic objectives (Grisé *et al.*, 2022). Russia's substantial investment in APT capabilities, their embedding within intelligence structures, and their persistent employment, suggest that Moscow expects strategic returns from these instruments. IPb doctrine also elevates information operations as tools for achieving state objectives across peacetime and wartime, implying that cyber capabilities should deliver value commensurate with their sophistication and integration.

What emerges from this theoretical framework are two testable hypotheses: H1) if APT groups function as instruments of state power maximisation, their operations should systematically align with Russian strategic objectives; this is evidenced through a target selection that reflects state priorities and timing coordinated with geopolitical context; and H2) given Russian investments in APTs, operations should achieve cumulative strategic effects, such as intelligence advantages, policy changes, or capability degradation, contrary to what offensive realism posits for cyberspace. The analysis tests these hypotheses through three case-studies.

3. Methodology

This study employs the "structured, focused" comparison method (George and Bennett, 2005). It is structured in that the researcher formulates standardised general questions derived from the research objective, which are then systematically applied to each case to guide data collection and enable a comparison. The method is focused in that it examines only those aspects of the historical cases that are directly relevant to testing the hypotheses.

The case-studies were selected according to criteria of i) high-confidence attribution and ii) a variation in operational context, to test if coordination and effect expectations hold across different competition phases. The first case-study is that of the 2015 Ukraine power grid attack, which occurred during armed conflict short of open war, representing the first successful cyberattack causing physical disruption of critical infrastructure. The second case-study is the 2020 SolarWinds supply chain compromise, a high-profile espionage operation targeting U.S. sanctions policy apparatus, representing peacetime strategic competition. The third case is the 2022 Ukraine war, examining APT deployment during active conventional warfare, and testing effects in an instance where "Bitskrieg" predictions were prominent (Brantly and Brantly, 2024).

The questions asked for each case are as follows: i) Which APT group conducted the operation, and to which Russian intelligence service are they affiliated? ii) What were Russia's strategic objectives during this period? iii) How did operation timing, targeting, and methods align with those objectives? iv) Which kind of effects were achieved? Questions i-iii) speak to H1. Question iv) tests H2 by distinguishing tactical effects (immediate disruption or intelligence collection) from strategic effects (sustained advantages, policy changes, and capability shifts) and identifying responses that constrain strategic gains.

The literature includes official Russian documents, articles by Russian military theorists, peer-reviewed literature, and technical reports from reputable cybersecurity firms. The research is also supported by reputable cybersecurity and news outlets when necessary.

One of the limitations of this study concerns assessing strategic effects, which require inferring causation from temporal correlation; this is addressed through convergent evidence across multiple sources.

4. Information Confrontation: Doctrine and Operational Ecosystem

The early sponsorship of "patriotic hackers" in cyber warfare in Estonia and Georgia represents the baseline experience of Russian use of cyberspace to achieve political and strategic objectives. Cyber warfare was employed in the first case to react to diminishing Russian influence in the Baltics (Giles, 2011; Haataja, 2017); while in the second, it was to assist in kinetic operations during conventional warfare (Giles, 2011; White, 2018). With the second achieving mixed results, the failures were attributed by Russian officials to the lack of professionalised "information troops", tasked with cyber operations and directed by a military command (Giles, 2011; Lysenko and Brooks, 2018). In 2011, this reasoning was enshrined within an official armed forces document, where a provision was included to position "own" forces and means to ensure Russian information security in the territory of other states (Ministry of Defense of the Russian Federation, 2011), and once more in 2014, within the official military doctrine, where the armed forces are called "to develop forces and means for *information confrontation*" (Security Council of the Russian Federation, 2014).

"Information confrontation" is defined by Grisé *et al.* as "the purposeful use of offensive or defensive informational means to achieve political, economic, military, and other objectives during peacetime, competition, and wartime" (Grisé *et al.*, 2022). "Informational means" are here intended in the Russian sense, merging cyber and information operations (Giles, 2016).

IPb in itself is stratified over a century of different practices, consisting of i) Soviet "active measures" (Abrams, 2016), framing political violence and destabilisation through any means as fundamental in achieving the survival and success of the State; and ii) the theory of "reflexive control" (Thomas, 2019), which entails the manipulation of an opponent to make them behave to the manipulator's advantage.

Among Russian military literature, IPb is reportedly split between "information-psychological warfare", targeting personnel and populations, employed continuously, and "information-technological warfare", targeting technical systems during armed conflicts (Kvachkov, 2004; Chekinov and Bogdanov, 2013). This distinction is reportedly adopted in Russian doctrine (Giles, 2016), though it is often blurred in practice.

IPb is implemented through a diverse range of actors, including loosely organised patriotic hackers, State-funded disinformation operations, and APT groups institutionalised within intelligence services. This reflects

the earlier argument from both Russian military theorists and official documentation for the creation and establishment of "forces and means" allocated for IPb. Given the scope of the research, the focus here is on mapping relevant APT groups.

These are primarily embedded within military and intelligence structures, as seen in Figure 1, a distinction hinting at a more specialised area of operations that directly reflects the role of the intelligence agency they belong to, with military intelligence units favouring cyberattacks involving system disruption, SVR-aligned units targeting foreign and diplomatic targets, and FSB-related groups focusing on espionage. However, it is worth noting that this is not always the case. For example, in the 2016 DNC hack, GRU-aligned FancyBear and SVR-aligned CozyBear were both found responsible for what was essentially a foreign espionage operation, typically under SVR purview, while unaware of one another. This form of overlap reduces the effectiveness of operations and increases the likelihood of detection (Melella, Ferazza and Mersinas, 2024).

At the same time, information warfare is enabled by APT groups only insofar as it is about targeted attacks, rather than massive disinformation efforts. A notable example is once again the 2016 DNC hack, where the intention was to destabilise the U.S. Presidential election by strategically releasing information acquired through a cyber operation (Bowen, 2021).

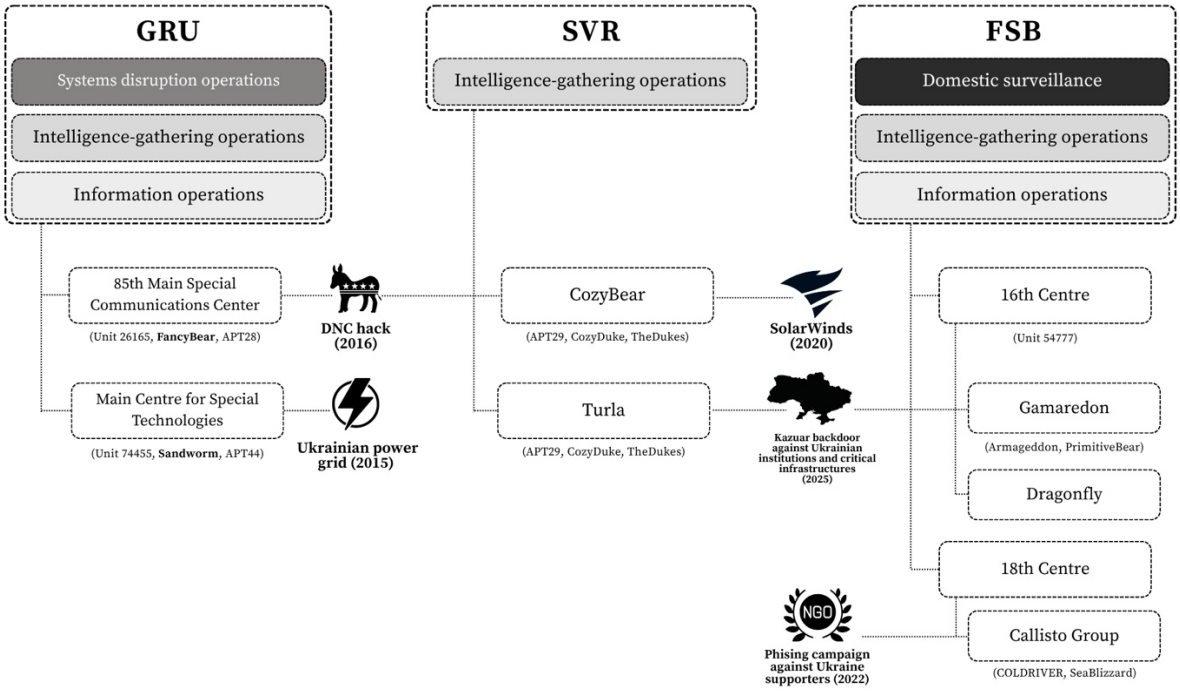


Figure 1: Authors mapping of the APT groups ecosystem under the Russian intelligence services. Key: intelligence services, their focus of operations, the APT groups attributed to them and some specific cyber operations (Cheravitch, 2021; Bowen, 2021, 2022; Wilde and Sherman, 2023; Melella, Ferazza and Mersinas, 2024; SSSCIPU, 2025)

5. Case-Studies

5.1 Ukraine Power Grid (2015)

Affecting 225,000 customers across three power-distribution territories, the December 23, 2015, cyberattack on the Ukrainian power grid is the first known case of a power outage caused by a cyber operation against critical infrastructure. The tools employed revealed high sophistication, with the attackers delivering BlackEnergy 3, a malware associated with GRU-affiliated Sandworm, through a phishing campaign (Hultquist, 2016; Lee, Assante and Conway, 2016; Greenberg, 2019). The strategic motive, as suggested by Baezner (2018), was to influence the population and the authorities, demonstrating the risks of escalation and resistance. This is evidenced through i) the hindering of crisis response capabilities by executing a telephone denial-of-service campaign (Hultquist, 2016); ii) the nature of a blackout at a time when no ground attack was taking place, as at the time the conflict between Ukraine and Russia was limited to few clashes, with the Second Minsk Accords still in place; iii) the timing (close to the Winter holidays).

Thus, the cyberattack aligns with Russian state objectives (H1), as the operation aimed to weaken Ukraine's resolve through infrastructure disruption. However, the attack's strategic impact was limited, contrary to what is stated in H2, as power was restored within hours (Kolodii, 2024), and no strategic objective in Ukraine was reached, as Kyiv continued to resist Russian pressure, even learning from this experience for the future 2022 conflict (Kolodii, 2024).

5.2 SolarWinds Attack (2020)

Between 2019 and 2020, SVR-aligned unit CozyBear infiltrated the supply chain of Orion, a data-managing software developed by SolarWinds. At the time, Orion was used by numerous entities worldwide, particularly in the U.S., where it was employed by key players such as Cisco, Microsoft, and at least twelve federal agencies (Devanny, Martin and Stevens, 2021). The malware spread through an update, resulting in a massive digital espionage campaign that affected over 18,000 SolarWinds clients (Willett, 2021).

Among the infected targets were also the U.S. Departments of Homeland Security and of the Treasury. While the exfiltrated information included counterintelligence information, Microsoft highlighted the access to data on sanctions via the Department of the Treasury as the primary objective of the operation (Menn and Bing, 2021).

The SolarWinds attack demonstrates coordination with state strategic objectives, supporting H1, as CozyBear targeted specific policy-relevant intelligence aligned with Russian priorities regarding U.S. sanctions policy. However, the operation provoked several reactions in the United States, resulting in heightened sanctions and a more robust cyber defence structure (Devanny, Martin and Stevens, 2021), limiting strategic gains regarding sanctions, despite the intelligence value of the exfiltrated data, which also included code, high-level correspondence and tools (Willett, 2021). These results contradict what was posited in H2, pointing to null strategic gains and only limited tactical ones.

5.3 APT Groups During the war in Ukraine (2022-)

During the 2022 invasion of Ukraine, Russian-sponsored APT groups subordinated to the GRU, FSB, and SVR were mobilised according to each intelligence service's distinct operational mandates and target sets.

GRU-linked units executed some of the most strategically timed attacks in the early phases of the war. Sandworm deployed AcidRain against KA-SAT satellite modems in February 2022, disrupting Ukrainian military communications at the onset of the kinetic operations (Boschetti, Falco and Gordon, 2022; Miron and Thornton, 2024; MITRE ATT&CK®, 2025). The attack on Viasat coincided with cyber operations targeting ISPs operating in the areas that were the main focuses of the invasion (NetBlocks, 2022). The impact was, however, offset by the redundancy of communication systems enjoyed by the Ukrainian military and civilian population across ISPs, mobile telephone providers, and secure radio communications (Brantly and Brantly, 2024).

Sandworm was later responsible for Industroyer2, an updated malware launched in April 2022, specifically designed to disrupt the industrial control systems of energy providers, with the potential to plunge over two million people into darkness (Wright, 2022); however, the attacks failed. These were synchronised with physical strikes on key logistical nodes in Kyiv, Vinnytsia, and Odesa (Kolodii, 2024), and were aimed at degrading physical infrastructure, causing confusion and eroding confidence, while also delaying decision-making in the field during offensive operations.

FancyBear, also GRU-affiliated, maintained a focus on credential harvesting, spear-phishing, and document exfiltration, targeting Ukrainian military and government institutions throughout 2022 and 2023. While not as technically destructive as Sandworm, FancyBear played a critical role in acquiring sensitive information, which was later exploited through information-psychological warfare (SSSCIPU, 2025).

FSB-controlled groups operated more diffusely across Ukraine's administrative and civilian structures, with the APT group Gamaredon being long focused on regional intelligence collection, with the number of phishing activities for espionage and intelligence-gathering purposes being high at the end of 2021 and the beginning of 2022, also scaling up after the beginning of the invasion, to the point where they became the most active APT group in Ukraine since the beginning of the war, with 829 attributed incidents (SSSCIPU, 2025).

Other FSB-linked APTs such as Callisto and Turla pursued credential theft and espionage campaigns against state institutions and civil society targets (CERT-UA, 2023).

What emerges from this brief analysis is that APT operations were deeply coordinated with the tactical and strategic objectives of the Russian Federation, rather than the groups acting as independent units, or "patriotic

hackers". This element supports H1 in proving the coordination of APT units with Moscow and the doctrine of IPb. However, while initial fears of a Russian "Bitskrieg" mounted before the beginning of the invasion (Brantly and Brantly, 2024), the experience shows that even the disruptive attacks that were successfully deployed had either temporary and limited impact; this is accounted by the Ukrainian defense capabilities as well as their lessons learnt throughout the previous periods of cyber warfare since 2014 (Kolodii, 2024). Therefore, H2 is not supported, with Russia being unable to deliver a cyber "kill-blow" to Ukraine.

6. Analysis and Conclusions

Table 1: Structured focused comparison of the case-studies

Questions	2015 Ukraine power grid	2020 SolarWinds	2022 Ukraine war
i) APT group and agency	Sandworm (GRU)	CozyBear (SVR)	Multiple: Sandworm, FancyBear (GRU); CozyBear, Turla (SVR); Gamaredon, Callisto (FSB)
ii) Context-specific Russian strategic objectives	To maintain pressure on Ukraine during the stalemate of the Minsk Accords, and showcase their capability to inflict damage without open war.	Gather intelligence on U.S. sanction policy; evade sanctions.	Degrade and disrupt Ukrainian C2 capabilities and critical infrastructure during the full-scale invasion of the country.
iii) coordination	Yes, to maximise psychological effects.	Yes.	Yes, with multiple instances of kinetic operations.
iv) Achieved tactical, strategic or response effects	Tactically, the operation resulted in a wide blackout. Strategically, the power was restored within hours, no policy change was registered and Ukraine hardened its cyber defences.	Tactically, important data was exfiltrated from U.S. institutions. Strategically, this resulted in a wider range of sanctions.	Tactically, disruptions only limitedly helped gains on the ground. Strategically, Ukrainian redundancies were able to offset the gains.

The case-studies showcase a consistent pattern of APT behaviour within Russian information confrontation doctrine: across all three instances, APT units systematically align with state objectives, rather than acting autonomously: the power grid attack targeted Ukrainian resolve at a critical time of competition between Kyiv and Moscow; CozyBear's infiltration of SolarWinds targeted the U.S. sanctions policy apparatus, reflecting Moscow's strategic priorities; finally, the 2022 campaigns demonstrated coordination between cyber operations and kinetic military actions. Thus, H1 is supported: APT groups function as institutionalised instruments of IPb, coordinated with state strategic objectives through their embedding within GRU, SVR, and FSB structures.

However, H2 reveals a gap between investments and outcomes. Despite Russia's substantial investment in APT capabilities, operations achieve tactical, rather than strategic effects, and limited ones at that (see Table 1). In the 2015 case-study, power was restored within hours, and Ukraine hardened its defences against future attacks, leading to the "pedagogy of cyber-war" reported by (Kolodii, 2024). In the case of SolarWinds, while intelligence exfiltrated, the discovery triggered heightened sanctions and expanded U.S. cyber defences, which offset the gains. In the wider 2022 campaign in Ukraine, initial "Bitskrieg" expectations were unmet, as Ukrainian resilience, communications redundancy, and allied support thwarted operations. Despite tactical differences across agencies, all APT groups faced similar strategic constraints when their operations prompted defensive adaptations or diplomatic retaliation.

From an offensive realist perspective, these findings reveal both the framework's explanatory power and its limitations. Offensive realism successfully explains why Russia organises and employs APT groups: as a revisionist power in anarchic cyberspace, Russia has institutionalised sophisticated cyber capabilities to maximise relative power through low-cost, deniable operations. The organisational embedding, target selection, and persistent employment align precisely with offensive realist logic about state behaviour in competitive domains. However, the expectation that sophisticated offensive tools yield cumulative strategic advantages proves limited. Rather, a more consistent pattern emerges: tactically successful operations trigger balancing responses, which then constrain the advantages gained during the operations. These constraints align with realist scholarship that remains sceptical of cyberattacks' "revolutionary" potential (Gartzke, 2013; Rid, 2013; Smeets, 2018), while also going further: even when operations coordinate with state objectives, employ sophisticated capabilities, and synchronise with conventional force, they achieve limited tactical effects without decisive strategic outcomes.

These findings clarify the role of APT groups in information confrontation. Russian-sponsored APTs remain sophisticated instruments that systematically pursue state objectives while achieving tactical rather than strategic effects. The gap between IPb doctrine's vision of cyber operations as tools of decisive power projection and their constrained operational reality reveals a significant gap: offensive realist imperatives drive Russia to maximise cyber capabilities and employ them persistently, yet the cyber domain characteristics, such as the adaptability of cyber defence, limit the strategic yield of even sophisticated operations. APT groups thus occupy a paradoxical role, remaining indispensable tactical instruments whose employment generates the very balancing responses that constrain their strategic value.

AI declaration: Grammarly was used for spelling/proofreading. All research content and analysis are entirely the authors' original work.

Ethics declaration: This research did not involve human participants or personal data and therefore did not require ethics approval.

References

Abrams, S. (2016) "Beyond Propaganda: Soviet Active Measures in Putin's Russia", *Connections: The Quarterly Journal*, 15(1), pp. 5–31.

Allen, P.D. and Gilbert, D.P. (2009) "The Information Sphere Domain Increasing Understanding and Cooperation", in C. Czosseck and K. Geers (eds) *The Virtual Battlefield: Perspectives on Cyber Warfare*. Tallin: CCDCOE, NATO Cooperative Cyber Defence Centre of Excellence.

Baezner, M. (2018) *Cyber and Information Warfare in the Ukrainian Conflict*. ETH Zurich.

Boschetti, N., Falco, G. and Gordon, N. (2022) "Space Cybersecurity Lessons Learned from The ViaSat Cyberattack", in. *AIAA Ascend 2022*, Las Vegas.

Bowen, A.S. (2021) *Russian Military Intelligence: Background and Issues for Congress*. R46616. Congressional Research Service.

Bowen, A.S. (2022) *Russian Cyber Units*. IF11718. Congressional Research Service.

Brantly, A.F. and Brantly, N.D. (2024) "The bitskrieg that was and wasn't: the military and intelligence implications of cyber operations during Russia's war on Ukraine", *Intelligence and National Security*, 39(3), pp. 475–495.

CERT-UA (2023) *Targeted attacks by Turla using CAPIBAR and KAZUAR malware*, *cert.gov.ua*.

Chekinov, S.G. and Bogdanov, S.A. (2013) 'The Nature and Content of a New Generation War', *Military Thought*.

Cheravitch, J. (2021) *The Role of Russia's Military in Information Confrontation*. Center for Naval Analyses.

Devanny, J., Martin, C. and Stevens, T. (2021) 'On the strategic consequences of digital espionage', *Journal of Cyber Policy*, 6(3), pp. 429–450.

Eriksson, J. and Newlove-Eriksson, L.M. (2021) 'Theorizing technology and international relations: prevailing perspectives and new horizons', in G. Giacomello, F.N. Moro, and M. Valigi (eds) *Technology and International Relations: The New Frontier in Global Power*. Edward Elgar Publishing.

Gartzke, E. (2013) "The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth", *International Security*, 38(2), pp. 41–73.

George, A.L. and Bennett, A. (2005) *Case studies and theory development in the social sciences*. Cambridge, Mass: MIT Press.

Giles, K. (2011) "'"Information Troops" – a Russian Cyber Command?", in. *3rd International Conference on Cyber Conflict*, Tallinn, Estonia: CCDCOE, NATO Cooperative Cyber Defence Centre of Excellence, pp. 45–60.

Giles, K. (2016) *Handbook of Russian Information Warfare*. Rome: NATO Defense College (Fellowship Monograph, 9).

Greenberg, A. (2019) *Sandworm: a new era of cyberwar and the hunt for the Kremlin's most dangerous hackers*. New York: Doubleday.

Grisé, M. *et al.* (2022) *Rivalry in the Information Sphere: Russian Conceptions of Information Confrontation*. RAND Corporation.

Haataja, S. (2017) "The 2007 cyber attacks against Estonia and international law on the use of force: an informational approach", *Law, Innovation and Technology*, 9(2), pp. 159–189.

Hultquist, J. (2016) *Sandworm Team and the Ukrainian Power Authority Attacks*. Mandiant.

Katagiri, N. (2024) "Advanced persistent threats and the "big four": State-sponsored hackers in China, Iran, Russia, and North Korea in 2003–2021", *Comparative Strategy*, 43(3), pp. 280–299.

Kello, L. (2017) *The Virtual Weapon and International Order*. New Haven: Yale University Press.

Kello, L. (2022) *Striking back: the end of peace in cyberspace - and how to restore it*. New Haven: Yale University Press.

Kolodii, R. (2024) "The Pedagogy of Cyber-War: Explaining Ukraine's Resilience Against Russian Cyber-Aggression", *Defense & Security Analysis*, 40(2), pp. 270–291.

Kvachkov, V. (2004) *Russia's Special Purpose Forces*. Voyennaya Literatura.

Lee, R., Assante, M. and Conway, T. (2016) *Analysis of the Cyber Attack on the Ukrainian Power Grid*. E-ISAC.

Lilly, B. and Cheravitch, J. (2020) "The Past, Present, and Future of Russia's Cyber Strategy and Forces", in. *2020 12th International Conference on Cyber Conflict (CyCon)*, IEEE, pp. 129–155.

Lindsay, J.R. (2013) "Stuxnet and the Limits of Cyber Warfare", *Security Studies*, 22(3), pp. 365–404.

Lynch III (2025) "Forward Persistence in Great Power Cyber Competition", *The Cyber Defense Review*, 9(3), pp. 81–103.
Lysenko, V. and Brooks, C. (2018) 'Russian Information Troops, Disinformation, and Democracy', *First Monday*.
Martino, L. (2018) "La quinta dimensione della conflittualità. L'ascesa del cyberspazio e i suoi effetti sulla politica internazionale", *Politica & Società*, (1), pp. 61–76.
Mearsheimer, J.J. (2001) *The Tragedy of Great Power Politics*. 1st edn. New York: Norton.
Melella, C., Ferazza, F. and Mersinas, K. (2024) 'Disjointed Cyber Warfare: Internal Conflicts among Russian Intelligence Agencies', *Applied Cybersecurity & Internet Governance*, 4(2).
Menn, J. and Bing, C. (2021) "Hackers of SolarWinds stole data on U.S. sanctions policy, intelligence probes", *Reuters*, 8 October.
Ministry of Defense of the Russian Federation (2011) "Conceptual Views Regarding the Activities of the Armed Forces of the Russian Federation in the Information Space".
Miron, M. and Thornton, R. (2024) "The Use of Cyber Tools by the Russian Military: Lessons from the War against Ukraine and a Warning for NATO?", *Applied Cybersecurity & Internet Governance*.
MITRE ATT&CK® (2025) *AcidRain, Software S1125*.
NetBlocks (2022) *Internet disruptions registered as Russia moves in on Ukraine - NetBlocks*.
Nye, J.S. (2011) "Diffusion and Cyberpower", in Nye, J. S., *The future of power*. 1. ed. New York: PublicAffairs.
Rid, T. (2013) *Cyber war will not take place*. Oxford ; New York: Oxford University Press.
Security Council of the Russian Federation (2014) "The Military Doctrine of the Russian Federation".
Smeets, M. (2018) "The Strategic Promise of Offensive Cyber Operations", *Strategic Studies Quarterly*, (Fall 2018), pp. 90–113.
Smeets, M. (2022) *No shortcuts: why states struggle to develop a military cyber-force*. London: Hurst & Company (Oxford scholarship online Political Science).
SSSCIPU (2025) *War and Cyber: Three Years of Struggle and Lessons for Global Security*. Kyiv: State Service of Special Communications and Information Protection of Ukraine.
Steffens, T. (2020) *Attribution of Advanced Persistent Threats: How to Identify the Actors Behind Cyber-Espionage*. Berlin, Heidelberg: Springer Berlin Heidelberg.
Thomas, T. (2019) *Russia's Reflexive Control Theory: Manipulating an Opponent to One's Advantage*. MITRE.
Voo, J. and Singh, V.V. (2025) *Russia's Information Confrontation Doctrine in Practice (2014–Present): Intent, Evolution and Implications*. International Institute for Strategic Studies, p. 36.
White, S.P. (2018) *Understanding Cyberwarfare: Lessons from the Russia-Georgia War*. Modern War Institute at West Point.
Wilde, G. and Sherman, J. (2023) *No Water's Edge: Russia's Information War and Regime Security*. Washington, DC: Carnegie Endowment for International Peace (Working Papers).
Willett, M. (2021) "Lessons of the SolarWinds Hack", *Survival*, 63(2), pp. 7–26.
Wright, R. (2022) *Industroyer2: How Ukraine avoided another blackout attack*, *TechTarget*.

Cybersecurity Issues and Solutions Within the South African Small and Medium-Sized Enterprises

Benediction Kitwa Kalombola and Tabisa Ncubukezi
Cape Peninsula University of Technology, Cape Town, South Africa

222689900@mycput.ac.za
ncubukezit@cput.ac.za

Abstract: This systematic literature review paper presents the main cybersecurity issues and their challenges within the Small and Medium-sized Enterprises (SMEs) in South Africa. The SME's play a very crucial role in the growth of the country's economy and Gross Domestic Product (GDP). However, following the global pandemic, the economy has become increasingly dependent on technology as a driver in different sectors. With the rapid adoption of technology, SMEs constantly face significant cybersecurity issues and challenges that regularly demand proactive and effective measures. These measures guard against exposure to sophisticated cyber-attacks. This study examines the state of cybersecurity within South African SMEs by assessing the effective use of cybersecurity measures, exploring the extent of their implementation, and identifying best practices to strengthen cyber resilience within these SMEs. To achieve this aim, a systematic literature review is conducted to examine high-quality peer-reviewed journals from 2021 to July 2025, providing more insight into the issues associated with cybersecurity in SMEs. Results highlight the vulnerability of SMEs to various cyber threats, including phishing, insider threats, and ransomware. The lack of awareness among employees, inadequate cybersecurity measures, limited resources, the shortage of professional experts, and the absence of effective measures specifically tailored to the needs of SMEs. Additionally, the mitigation strategies emphasize the adoption of robust security measures that are specifically tailored to the needs of SMEs. To enhance cyber resilience, this study also recommends the use of cybersecurity measures tailored to SMEs, encouraging employee education training programs, and the creation of reliable strategies.

Keywords: Best practices, Cybersecurity, Cybersecurity challenges, Cybersecurity issues, Cybersecurity measures, Cyber threats, Small businesses, Systematic literature review

1. Introduction

In South Africa, small and medium-sized enterprises (SMEs) play a crucial role in driving economic growth; however, SMEs have become increasingly vulnerable to cyber threats with the rapid adoption of technology. Despite the benefits associated with digital transformation in improving operational activities, it has also introduced new threats, leaving SMEs exposed to threats such as data breaches, ransomware, and phishing, mainly due to the lack of awareness, limited resources, and ineffective measures (Fakir, 2024; Saeed *et al.*, 2023; Ncubukezi, 2022a). Additionally, because SMEs often rely on outdated security measures and ineffective strategies, they are frequently left vulnerable to threats, particularly in the financial and healthcare sectors (Mahlangu, 2023; Ncubukezi, 2022b).

SMEs encounter issues and challenges in adopting effective cybersecurity measures due to inadequate management and regulations, the absence of skilled professionals, and resources (Shingange, 2022). According to Phahlamohlaka *et al.*, (2022), there is a gap between the practical implementations and available frameworks, including the Zero Trust Architecture (ZTA), National Cybersecurity Policy Framework (NCPF) and the National Institute of Standards and Technology (NIST), which is further enlarged by SMEs unawareness and inability to adopt advanced tools such as multi-factor authentication, encryption, and artificial intelligence (AI) and machine learning (ML) for early threat detection (Admass *et al.*, 2024). Considering that SMEs employ almost 60% of the working population and constitute around 90% of registered businesses in the nation, the impact of cyber threats on the financial institutions is enormous in South Africa. Because human mistakes are the primary reasons that trigger data breaches, Reshmi (2021) argues that this is largely due to the lack of adequate cybersecurity training among employees.

Cybersecurity has become a critical concern for South African SMEs. Businesses are increasingly exposed to sophisticated threats and face challenges in protecting their data (Ncubukezi & Mwansa, 2021). As technology adoption advances, it has become increasingly crucial for SMEs to comprehend the complexities of cybersecurity threats in order to survive and thrive (Benjamin *et al.*, 2024). The problem stems from the lack of effective security measures and the reliance on outdated systems, which are now less effective in mitigating sophisticated threats. The emergence of advanced cyber threats causes several damages within SMEs, including operational disruption, loss of reputation, and economic loss, making it difficult for these businesses to protect the sensitive data they handle (Dave *et al.*, 2023; Mahlangu, 2023).

While several studies addressed the key importance of affordable and customized security measures in which SMEs' real-world challenges are considered, there is still a gap regarding the specific issues and challenges encountered by SMEs along with practical and tailored solutions (Benjamin *et al.*, 2024). The lack of modern and tailored security measures leaves the SMEs vulnerable and constantly exposed to sophisticated cyber threats. Without addressing this, both SMEs and financial institutions will remain vulnerable to cyber threats and suffer the associated consequences. Hence, the study intends to examine the major issues and challenges, exploring the extent of implementation of existing security measures, and providing best practices and effective measures to enhance cyber resilience within SMEs.

To achieve the aim, this work will carry out the following objectives:

- To investigate the major cybersecurity issues experienced by the SMEs in South Africa.
- To explore the extent of implementation of existing measures.
- To investigate the best practices for effective security measures.

2. Literature Review

Several studies have examined the ever-evolving field of cybersecurity. Despite this, a thorough examination of this study is imperative, considering the specific obstacles that most Small and Medium-sized Enterprises (SMEs) face in South Africa (Fakir, 2024). While studies provide recommendations about security threats, the challenges experienced by SMEs remain largely unaddressed (Benjamin *et al.*, 2024). Hence, there is a need for local studies that strongly focus on the specific challenges faced by these businesses. SMEs are becoming more vulnerable to sophisticated cyber threats. However, studies have revealed the challenges these businesses frequently experience in adopting effective security measures, highlighting the need for more advanced strategies tailored to them (Saeed *et al.*, 2023). However, there is still a gap between awareness, current strategies and the actual implementation and regulations (Phahlamohlaka *et al*., 2022). The gradual growth of the technology-driven economy has brought both benefits and emerging threats to SMEs, despite their significant contribution to the economy. Due to digital transformation, these businesses face several issues and challenges in protecting their digital assets, primarily due to a lack of effective security measures and resource constraints (Fakir, 2024), leaving them vulnerable and defenceless against sophisticated threats (Dave *et al.*, 2023). While more advanced frameworks, such as NIST and Zero Trust Architecture (ZTA), exist, their adoption remains limited due to the lack of skilled professionals, financial support, and proper policies required for their implementation (Bokan & Santos, 2021; Lokare *et al.*, 2025). Despite growing awareness of cybersecurity risks, a gap remains between knowledge, strategies, and actual practical actions, especially in resource-constrained environments, including SMEs (Lokare et al., 2025; Admass et al., 2024).

2.1 Summary of Existing Literature

Benjamin et al. (2024) highlight that SMEs are the primary target of cyber threats due to the lack of effective measures and skilled professionals while adopting technology. SMEs are frequently exposed to cyber threats due to inadequate security measures, underscoring the need for more comprehensive and tailored security measures to enhance the protection of their digital data and ensure smooth business operations. Mahlangu (2023) examines the growing concern that emerging ransomware attacks have become a critical concern for financial institutions. The leadership, as well as the behaviour of people, plays a vital role in strengthening endurance within organizations (Mahlangu, 2023), drawing attention to the fact that defense against ransomware is unlikely to be effectively eradicated with technical implementations only. In consideration of risks and challenges faced by financial institutions and businesses, Mahlangu (2023) advocates for measures specifically tailored to these institutions (Mahlangu, 2023). Admass et al. (2024) investigate the role of Machine Learning (ML) and Artificial Intelligence (AI) to strengthen cybersecurity. Admass et al. (2024) highlight the difficulties resulting from insufficient knowledge and flexibility commonly encountered by the majority of SMEs, despite the fact that potential solutions could arise from technological advances such as these. In most cases, SMEs are unable to guarantee effective security due to universal security measures that fail to meet the specific needs of SMEs (Admass et al., 2024). Hence, there is an urgent need for specifically tailored measures to address the challenges faced by SMEs. The study by Saeed *et al.* (2023) highlights that while the adoption of technology may contribute to business growth, it also exposes frequent SMEs to cyber threats. In addition to fostering risk-based security strategies, Saeed *et al.* (2023) underline a pressing need for proactive measures tailored to SMEs, while also suggesting the adoption of flexible and innovative strategies. The study by Fakir (2024) examines the most significant challenges associated with adopting cybersecurity. Because almost every organization operates with various activities and management structures, the difficulties

also vary from one to another organization (Fakir, 2024). Hence, there is a need for more adaptive approaches tailored to the specific needs and challenges faced by SMEs, due to the ineffectiveness of the majority of generic solutions. Additionally, there is a need for continued investigations to closely monitor the advancement of cybersecurity requirements within SMEs (Fakir, 2024).

Shingange (2022) focuses on the lack of connection that exists between the design of measures and their implementation in practice, while investigating the challenges associated with cybersecurity laws and regulations in South Africa. Shingange (2022) highlights the ongoing absence of communication between private industries and governments, which weakens the overall effectiveness of cybersecurity operations. Due to a failure to enforce current laws and regulations, there is a need to critically evaluate and strengthen them (Shingange, 2022).To ensure successful progression, there is a need for greater coordination along with further visible enforcement. In 2021, Reshmi investigated ransomware attacks and their prevention mechanisms. While more advanced technologies like Artificial Intelligence (AI) and Machine Learning (ML) have the potential to identify threats early, their application remains limited within SMEs. This study suggests a wider integration of AI/ML to enhance security against ransomware.

The CENSOR model, as outlined by Tsiodra et al. (2023), is a risk-based model for cybersecurity decision-making. It is specifically beneficial to the SMEs as it guides them in prioritizing their limited resources by evaluating threats according to their significance. The authors also provide an illustration demonstrating how SMEs can apply this model to allocate resources more effectively. Tsiodra et al. (2023) explore the relationship between the magnitude of cyber threats that SMEs encounter and the level of investment that SMEs make. Hence, the critical importance of establishing an appropriate balance between budgeting and maintaining effective security. In the same way, Hermanus (2023) examined the vulnerabilities associated with information security within the financial environment in Cape Town. As reported by the participants in the interviews, organizations continue to face serious problems towards the full adoption such as insufficient resources, the shortage of professional experts, and a poor organized response strategy, despite the fact that some enterprises started putting into effect security awareness training, regulations to the protection of information, intrusion detection systems as well as authentication and authorization systems (Hermanus (2023). Hence, this reveals that some enterprises are becoming aware and starting to take effective action; however, the complete adoption is constrained by operational issues.

In 2023, Musa investigated the cybersecurity issues faced by SMEs from a systems engineering perspective. This study highlights the lack of ongoing monitoring, employee training, and incident response, as the majority of SMEs rely on basic strategies such as antivirus, and password policies. The absence of skilled professionals, resource shortages, and the need for tailored security measures are a few barriers preventing the adoption of cybersecurity. In their review of literature, Alveera & Sukira (2024) delt with cybersecurity within the SMEs, highlighting their susceptibility to cyber-attacks from malicious computer programs and ransomware, regardless of the importance of these SMEs to the growth of the economy within the country, which is mainly the consequence of several factors, including the limited financial resources, inadequate or outdated infrastructures, and regular misperception of the complexity of cybersecurity. Additionally, Alveera & Sukira (2024) highlight the fact that SMEs are often given little attention compared to large corporations, which are mainly prioritized by government initiatives for cybersecurity. This study recommends the adoption of rigorous security mechanisms specifically tailored to the needs of SMEs, conducting risk assessments, and educating team members about cyber threats to ensure robust cybersecurity for SMEs (Alveera & Sukira, 2024).

In 2024, Mugwagwa *et al.* highlight that due to poor awareness, resource constraints, and the absence of skilled professionals, SMEs remain vulnerable to complex threats like insider threats, ransomware, and phishing, highlighting the need for effective malware detection, multi-factor authentication, encryption, along regular employee education to effectively enhance cyber resilience within SMEs. In summary, while there is a growing awareness of cybersecurity, a significant gap remains between awareness, strategic plans, and actual implementation. While cutting-edge technologies like AI/ML can enhance cyber resilience within SMEs, their adoption remains limited due to financial constraints, resource shortages, and a lack of skilled professionals, which are primary issues faced by SMEs in South Africa, preventing them from implementing effective cybersecurity measures.

3. Methodology

A comprehensive systematic literature review is employed in this study to investigate the state of cybersecurity in South African SMEs, with a strong focus on the major issues and challenges, as well as the

extent of implementation of existing security measures. The objective is to provide best practices for enhancing their resilience to cybersecurity. The systematic literature review is adopted in this study along with the analysis of the contents to discover, critically review, and prioritize key research that directly aligns with the previously established questions (using a systematic literature review) and come up with commonly identified issues and challenges, as well as important and repeated solutions and recommendations (content analysis).

3.1 Selection Criteria

To ensure a comprehensive examination of the available literature, the data used were collected from various important databases, including Google Scholar, IEEE Xplore, ScienceDirect, and MDPI. Additionally, South Africa's sources were particularly included to enable a more objective and in-depth examination of the challenges associated with cybersecurity in the South African SMEs.

3.2 Search Strategy

To gather relevant sources of information, a systematic search was conducted. The search was conducted using keywords and phrases including "Cybersecurity" AND "South Africa" AND "SMEs," “Cybersecurity in SMEs,” “Cybersecurity,” “Cybersecurity challenges,” “Information security in SMEs,” Online security,” “Cybersecurity issues,” “Digital security,” “Cyber issues,” “Cyber resilience strategies for SMEs,” and “Data breaches.”

3.3 Prisma Flow Diagram

This study adheres to the PRISMA principle to ensure clarity in the selection of literature, with the aim of identifying, examining, and selecting the most suitable studies for this research. In accordance with predefined inclusion and exclusion criteria, the overall approach consisted of several narrowing phases. For this reason, the following Figure 1 below presents an illustration of the method by which information progresses throughout the different stages of this process of review:

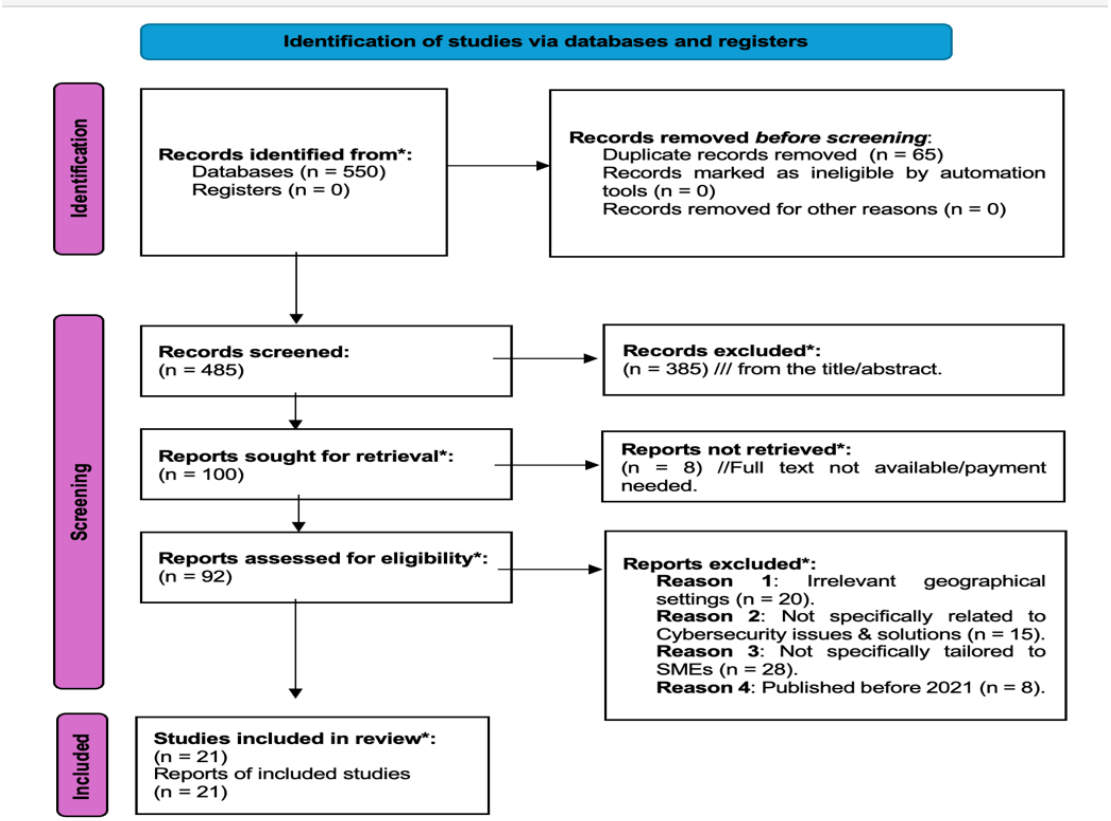


Figure 1: Prisma Flow Diagram

3.4 Inclusion and Exclusion Criteria for Relevant Literature

To gather more accurate and relevant studies, the inclusion and exclusion criteria were defined in this study. Therefore, this research examined various studies that discuss cybersecurity issues and challenges within SMEs, with a focus on studies that emphasize the development of security measures specifically tailored to the needs of SMEs. Additionally, only English articles published between year 2021 and July 2025 were considered to collect the most recent information. On the other hand, excluded studies were those that did not focus on SMEs' issues, challenges, and their needs. Additionally, articles published before 2021 were excluded to ensure that this investigation aligns successfully with the latest information and developments. By following these guidelines, the review was focused on the predefined objectives.

3.5 Data Analysis

To identify important themes, challenges, and technological innovations associated with cybersecurity for SMEs, the content analysis was conducted on the selected literature, aiming to identify issues and challenges encountered by SMEs and potential innovative solutions to enhance their cyber resilience. Finally, the results were combined to provide actionable recommendations to enhance cyber resilience within South African SMEs.

4. Results and Discussions

A total of four interrelated points were revealed from the review of the literature: the different kinds of threats commonly associated with cybersecurity that SMEs experience, the issues preventing them from adopting and using effective security measures, and the extent to which these measures are successfully implemented. SMEs have begun to understand the necessity of introducing effective cybersecurity measures, as the digital economy resulting from technological advancements continues to evolve. The use of these effective cybersecurity measures has become crucial for ensuring the protection of the digital assets they contain while maintaining business operations. The goal of this systematic literature review was to explore the primary issues, challenges, and potential solutions related to the effective implementation of cybersecurity in South African SMEs. Categorized into four key themes, the results of this research indicate that while threats to cybersecurity are gradually being recognized, the potential to effectively establish robust and effective countermeasures remains remarkable. In this section, the findings will be discussed along with their connection to the research objectives and questions.

4.1 Common Cybersecurity Threats Experienced by SMEs in South Africa

SMEs remain vulnerable to various cyber threats identified in the literature, including poor password policies, ransomware attacks, phishing, and insider threats (Mugwagwa et al., 2024; Reshmi, 2021). For businesses and financial institutions, ransomware poses a significant concern due to its potential to result in financial loss and business disruption (Mahlangu, 2023). SMEs have become vulnerable to various threats because of digital transformation, despite the benefits it has brought. Despite the growing awareness of cybersecurity among SMEs, the results of the study indicate that they frequently experience various issues and challenges that restrict them from effectively making use of all necessary measures, including the lack of skilled professionals, poorly structured organizations, outdated systems, and budget constraints (Hermanus, 2023; Musa, 2023; Benjamin *et al*., 2024). Issues with policies and regulations are also quite important. According to Shingange (2022), an inefficient approach to cybersecurity governance has resulted from the lack of strong enforcement and inadequate interaction between the public and commercial sectors. It is common for many SMEs to operate without clearly established cybersecurity policies and regulations, instead relying on improvised solutions, which highlights that cybersecurity is a problem that extends beyond technology and has a profound impact on the way organizations and their systems interact. Given this, there is a need for robust organizational support, effective leadership, and thoughtful policy interventions.

Additionally, Tetteh (2024) emphasizes that more structured threat detection, combined with incident reporting, is essential for achieving more effective cybersecurity in SMEs. According to Tetteh (2024), cyber resilience across the nation has become weaker due to the inability of SMEs to report security breaches, which limits the opportunities for education across the industry. Therefore, there is a need for simplified frameworks that allow SMEs to effectively report incidents and assess risks in a more adaptive manner.

4.2 Challenges Associated with the Effective Use of Cybersecurity Measures

The gap between knowledge and practical actions remains large. The reliance on obsolete systems, financial constraints, a lack of skilled professionals, and the need for tailored security measures are the primary

obstacles (Hermanus, 2023; Musa, 2023; Benjamin et al., 2024). Moreover, further challenges include the breakdown of collaboration between the private and public sectors, inconsistent directions for implementation, and a failure to enforce regulations, as well as inadequate management and weak internal processes (Shingange, 2022; Fakir, 2024). Additionally, the failure of SMEs to report cyber incidents significantly impacts the general and national response efforts. This highlights an ongoing issue with inadequate reporting and the lack of a structured prioritization framework (Tetteh, 2024). By utilising a practical model introduced in this study, Small and medium-sized enterprises are enabled to focus their resources on addressing the most urgent issues and challenges, which in turn enhances their overall resilience.

Although small and medium-sized enterprises (SMEs) are becoming aware of cybersecurity risks and challenges they experience, the review shows that the efforts they make to put in place practical and effective preventive measures are still at the basic level, which include antivirus software, constrained awareness education, and password regulations, while more advanced and sophisticated tools such as ongoing monitoring, systems for intrusion detection, and the Zero Trust Architecture (ZTA) remain rarely, or not used (Hermanus, 2023; Musa, 2023; Erdogan *et al*., 2023; Lokare *et al*., 2025). This circumstance illustrates a clear disconnect between awareness and action. Small and medium-sized enterprises (SMEs) often struggle to acquire cutting-edge technologies, primarily due to financial constraints and limited resources. Additionally, their ability to maintain or enhance complicated systems is limited due to their lack of technical expertise. In many cases, the use of cybersecurity is generally more of a response to specific incidents than an anticipatory measure, as highlighted in section 4.2 "Long-term Strategy," where businesses commonly wait for incidents to occur before taking action.

4.3 Extent of Implementation of Cybersecurity Measures

According to the reviewed articles, many SMEs are using basic security precautions, ranging from password rules, limited awareness training for employees, and anti-virus programs, while preventive and more sophisticated strategies exist, but continue to be left behind (Hermanus, 2023; Musa, 2023; Erdogan et al., 2023; Lokare et al., 2025). Additionally, certain organizations possess disaster recovery plans, but these plans often lack clear procedures. Due to their high costs and complexity, emerging technologies such as threat identification based on intelligence and machine learning (AI/ML), intrusion detection systems, and ongoing monitoring are rarely adopted (Admass et al., 2024; Lokare et al., 2025).

A variety of effective approaches for enhancing SMEs' resilience to cybersecurity threats are highlighted in this research, pointing out the adoption of risk-based strategies to enable SMEs to focus their resources on the greatest threats (Tsiodra *et al.*, 2023; Tetteh, 2024) and the necessity for international models, including the NIST framework for cybersecurity and the ZTA, to be adjusted for resource-constrained environments (Fakir, 2024; Bokan & Santos, 2021). Moreover, studies underlining the necessity for ongoing staff education and training which is indispensable for preventing errors among employees (Mugwagwa *et al.*, 2024; Benjamin *et al.*, 2024) while also promoting encryption, multi-factor authentication, and the need to identify threats at their early stage by integrating more advanced technologies such as AI/ML (Admass *et al.*, 2024; Reshmi, 2021). Additionally, studies promote collaboration between the private and public sectors to offer SMEs adequate funding and support (Shingange, 2022).

4.3.1 Best practices and recommendations

Studies suggest that to reconcile the divide between knowledge and practical action, it is necessary to tailor cybersecurity approaches to the needs of SMEs, adapting frameworks like Zero Trust Architecture (ZTA) and NIST to resource-constrained environments (Fakir, 2024; Bokan & Santos, 2021). To reduce human errors, a continuous educational program becomes crucial (Mugwagwa et al., 2024; Benjamin et al., 2024), as well as risk-based approaches for effective resource allocation (Tsiodra et al., 2023). Moreover, advanced measures such as AI/ML, multi-factor authentication, and encryption can enhance security through early threat detection (Admass et al., 2024; Reshmi, 2021). Additionally, studies suggest that the government should strengthen collaboration between the private and public sectors, increase policy monitoring, and provide more detailed recommendations tailored to the benefit of SMEs.

The need for more effective and robust cybersecurity measures has become paramount, particularly within SMEs, which continue to remain more vulnerable and are constantly exposed to cyber threats as a result of being primary targets. Thus, the following recommendations resulting from the results and conclusions drawn from this research will enable SMEs to further strengthen their resilience and utilization of cybersecurity in South Africa:

- **Promote awareness on cybersecurity**: To effectively handle human-related vulnerabilities while improving the culture of companies, there is a need for regular awareness and educational campaigns about cybersecurity.
- **Adapt frameworks tailored to SMEs' needs:** A necessity for both policymakers as well as experts in cybersecurity to design more reasonable security measures that adapt to the specific business processes, size, and the available resources of SMEs.
- **Promote the utilization of risk-based solutions**: As recommended by Tetteh (2024) and Tsiodra *et al.* (2023), SMEs could first detect and address high-risk spots, concentrating the insufficient resources they have on the biggest threats to cybersecurity.
- **Foster cooperation**: To ensure long-term viability within the SMEs, the findings of this research suggest collaboration between the private sector, government agencies, and industries in order to offer more affordable accessibility to support, financial resources, and equipment.
- **Promote the reporting of security incidents**: To gain more insight from cyber disasters and effectively strengthen cyber resilience within the nation, this study recommends the start-up of more secure and accessible ways to report incidents.

Table 1: Summary of cybersecurity threats, challenges, and practices in SMEs

Theme	Key Findings
Common cybersecurity threats	In S.A, SMEs are constantly confronted with a range of threats such as ransomware, malware, data breaches, and phishing, which frequently arise from weak credentials, lack of proper monitoring, and outdated systems.
Challenges to implementation	Primary obstacles consist of the lack of skilled cybersecurity professionals, poor enforcement of regulations, a general lack of awareness, lack of financial resources, and the struggle to convert national policies (e.g. NCPF) into effective action.
Extent of cybersecurity implementation	Despite the growing awareness about the importance of cybersecurity, implementation remains low. SMEs frequently rely on basic security measures such as Antivirus and firewalls, while neglecting more advanced strategies such as AI-based threat detection, well-structured risk management frameworks, and the MFA.
Best practices and recommendations	To enhance cybersecurity, it is essential to engage in ongoing training, refining password policies, using budget-friendly frameworks, using the CENSOR for resource allocation, promoting cooperation between governments, SMEs, and private sectors cybersecurity providers, and setting priorities based on risks.

5. Conclusion and Future Work

This study aimed to investigate issues and challenges associated with cybersecurity experienced by SMEs in South Africa, exploring the extent of implementation of existing security measures and providing potential best practices. The research reveals a poor state of cybersecurity within SMEs regardless of the growing awareness of cybersecurity threats. These businesses face ongoing challenges in adopting cutting-edge technological solutions due to a lack of employee awareness, insufficient resources, and a shortage of skilled professionals. Moreover, the practical integration of advanced security frameworks, such as ZTA and NIST, remains limited due to their high implementation cost, despite the potential they offer, which explains the reliance of SMEs on basic security strategies, including antivirus software and password policies. Additionally, the results suggest that SMEs often opt for a reactive plan of action, prioritizing a response to incidents over proactive preparation. However, cybersecurity should be considered both a strategic and an administrative issue, rather than a technical one only, as it is an essential way to handle all these issues. For South African SMEs to enhance their cyber resilience, it is necessary to integrate cybersecurity into their organizational culture. For this reason, the ultimate necessity is for security measures that are more affordable, more tailored to the specific needs of SMEs, capable of accommodating issues associated with their activities and resource constraints, thus strengthening the resilience of SMEs in South Africa to cybersecurity threats.

For future work, - further studies could be conducted on the improvement of small and medium-sized enterprises' (SMEs) potential towards threat detection and response in real-time, by incorporating algorithms as well as data structures within a programmed information security system, which could provide more

understanding on how the usage of these approaches could be improved to be beneficial in environments with restricted resources.

AI declaration: For the development of this work, no AI tools were used.

Ethics declaration: Since the study adopts a systematic review of previous literatures which focus on secondary data publicly made available without direct human interaction or raw data collection, no ethical approval was required. However, in order to preserve academic integrity and avoid plagiarism, all sources have properly been referenced.

References

Admass, W., Munaye, Y. & Diro, A. 2024. Cyber security: State of the art, challenges and future directions. Amsterdam: Elsevier BV.

Alveera, T. & Sukira, M. 2024. Cybersecurity challenges and resilience strategies in South African SMEs. Cape Town: UCT Press.

Benjamin, L., Adegbola, A., Amajuoyi, P., Adegbola, M. & Adeusi, K. 2024. Digital transformation in SMEs: Identifying cybersecurity risks and developing effective mitigation strategies. Vol. 19. GSC.

Bokan, B. & Santos, J. 2021. Managing cybersecurity risk using threat-based methodology for evaluation of cybersecurity architectures. New York: IEEE.

Dave, D., Sawhney, G., Aggarwal, P., Silswal, N. & Khut, D. 2023. The new frontier of cybersecurity: Emerging threats and innovations. Toba: IEEE.

Erdogan, T., Rahman, M. & Kwon, H. 2023. Assessing cybersecurity maturity among small and medium-sized enterprises. New York: IEEE.

Fakir, J. 2024. An investigation of cybersecurity implementation challenges among South African SMEs. Master's dissertation. Johannesburg: University of the Witwatersrand.

Hermanus, K. 2023. Information security vulnerabilities in financial institutions in Cape Town. Master's dissertation. Cape Town: University of Cape Town.

Lokare, A., Bankar, S. & Mhaske, P. 2025. Integrating cybersecurity frameworks into IT security: A comprehensive analysis of threat mitigation strategies and adaptive technologies. arXiv preprint. Available at: https://doi.org/10.48550/arXiv.2502.00651. (Accessed : 15 August 2025).

Mahlangu, N. 2023. Strategies to mitigate ransomware-related cyber-attacks in South African financial institutions. Master's dissertation. Johannesburg: University of the Witwatersrand.

Mugwagwa, T., Phiri, N. & Dube, C. 2024. Cybersecurity issues in South African SMEs: Risks and solutions. Pretoria: University of South Africa Press.

Musa, L. 2023. A systems engineering approach to cybersecurity challenges in South African SMEs. Master's dissertation. Durban: University of KwaZulu-Natal.

Ncubukezi, T. and Mwansa, L., 2021. Best practices used by businesses to maintain good cyber hygiene during Covid19 pandemic. *Journal of Internet Technology and Secured Transactions*, *9*(1), pp.714-721.

Ncubukezi, T., 2022a. Human errors: A cybersecurity concern and the weakest link to small businesses. In Proceedings of the 17th *International Conference on Information Warfare and Security* (p. 395).

Ncubukezi, T., 2022b. Impact of information security threats on small businesses during the COVID-19 pandemic. In *European Conference on Cyber Warfare and Security*, 21(1), pp. 401-410.

Phahlamohlaka, J., Theron, J. & Aschmann, M. 2022. National Cybersecurity Implementation in South Africa: The Conundrum Question. Vol. 21. Pretoria: ArmisteadTec.

Pieterse, H. 2021. The cyber threat landscape in South Africa: A 10-year review. Pretoria: LINK.

Reshmi, T. 2021. Information security breaches due to ransomware attacks – a systematic literature review. Amsterdam: Elsevier Ltd.

Saeed, S., Altamimi, S., Alkayyal, N., Alshehri, E. & Alabbad, D. 2023. Digital transformation and cybersecurity challenges for business resilience: Issues and recommendations. Basel: MDPI.

Shingange, J. 2022. Problematizing the South African cybersecurity policy landscape. Master's dissertation. Stellenbosch: Stellenbosch University.

Tetteh, A.K., 2024. Cybersecurity needs for SMEs. Issues in Information Systems, 25(2), pp.33–42. IACIS.

Tsiodra, M., Panda, S., Chronopoulos, M. & Panaousis, E. 2023. Cyber risk assessment and optimization: A small business case study. Vol. 1. New Jersey: IEEE.

Venkatasubramanian, D., Goyal, S., Ezhilarasan, G., Sharma, Y., Vijayalakshmi, V. & Garg, P. 2024. Evaluating the effectiveness of multi-factor authentication for preventing cyber attacks. Kamand: IEEE.

AI-Augmented Proactive Cyber-Detection and Mitigation of Cybersecurity Threats in the Banking Sector

Prince Rotondwa Mulea and Dewald Blaauw
Department of Information Science, Stellenbosch University, South Africa

Princeroem@gmail.com
Dnblaauw@sun.ac.za

Abstract: The digital transformation of the financial services sector, accelerated by the emergence of neobanks and advanced online platforms, has markedly increased its exposure to sophisticated cyberthreats. High-profile incidents, such as coordinated attacks on financial institutions in Iraq, have demonstrated the severe operational, economic, and reputational consequences that can arise from delayed threat detection and inadequate mitigation. Traditional cybersecurity measures, including firewalls, antivirus software, and signature-based intrusion detection systems, remain constrained by their dependence on known attack signatures, thereby leaving financial networks susceptible to zero-day exploits, AI-driven intrusions, and complex multi-vector threats. This study proposes and evaluates a supervised machine learning intrusion detection and prevention model aimed at proactively securing financial networks at a network level. To simulate realistic network conditions and generate representative traffic data, a banking environment was constructed using GNS3. To address class imbalance within the dataset, the Synthetic Minority Oversampling Technique (SMOTE) was employed, thereby improving the detection of minority-class attack instances. Several machine learning algorithms, including Support Vector Machine, Multi-Layer Perceptron Neural Network, and Long Short-Term Memory, were assessed using key performance metrics to determine their effectiveness. The Decision Tree model demonstrated superior performance, achieving an accuracy rate of 99.98%, perfect precision and recall, zero false positives, and only thirteen false negatives. These results underscore its capacity to deliver highly accurate, real-time threat detection while minimising operational disruptions caused by false alarms. Its transparent decision-making process enhances explainability, supports regulatory compliance, and fosters institutional trust, factors that are critical in financial cybersecurity. The findings validate the viability of interpretable, high-performance machine learning models for the real-time detection and mitigation of advanced cyberthreats, including Distributed Denial-of-Service (DDoS) attack patterns. Future research should prioritise scaling the simulation framework to encompass more complex financial network topologies, integrating adaptive online learning capabilities, and incorporating explainable artificial intelligence (XAI) techniques to investigate whether enhanced model interpretability improves threat detection accuracy and analyst response times.

Keywords: Machine learning, Cybersecurity, Cyberattacks, Financial networks, Banking sector

1. Introduction

The universal banking model, encompassing institutions that provide a wide range of financial services beyond traditional lending and payments, has historically shaped the global financial sector. Over time, this model has evolved through digital-first strategies, advanced data capabilities, and new technologies that enhance service delivery and competitiveness. Traditionally, high entry barriers limited competition, reinforcing the dominance of universal banks. However, the rise of neobanks, also referred to as digital banks, lowered these barriers and disrupted the sector's competitive landscape. Unlike the traditional model, modern banking now emphasises customer engagement, personalisation, and agile core systems, shifting priorities from sales and services to consumer experience.

While the digitisation and digitalisation of the financial sector, along with the integration of neobanks, have positively impacted the sector's productivity and operations, these technological advancements have also heightened cybersecurity threats within the financial industry. Beyond traditional security risks such as financial information breaches, cyberattacks now pose systemic challenges to financial infrastructure. These cyberattacks, which are the primary focus of this study, threaten not only banking stability but also national security and economic resilience.

The increasing frequency and sophistication of cyberattacks underscore the pressing need for solutions that not only respond to incidents post-occurrence but also proactively detect and mitigate malicious activities in real time, ensuring that threats such as large-scale fraud attempts are intercepted at their earliest stages. Consequently, effective mitigation strategies must operate in tandem with proactive detection mechanisms, establishing a multi-layered defence system capable of addressing both known and emerging threats. The mitigation of these attacks involves the utilisation of artificial intelligence (AI) tools and traditional cybersecurity measures.

This study proposes a potential solution for safeguarding the financial sector at a network level against cyberthreats and cybercriminals through the utilisation of AI tools. However, the accelerating speed and increasing complexity of contemporary threats have exposed the limitations of traditional cybersecurity systems, thereby underscoring the imperative for intelligent, adaptive models that can learn from and respond to evolving attack patterns.

2. Literature Review

The literature review examines cybersecurity within the banking sector, highlighting the integration of AI and ML as critical tools for enhancing security.

2.1 Cybersecurity and the Presence of AI in Cybersecurity of the Banking Sector

Technological developments in computational capacity and real-time communication have financial operations, enabling intelligent data management and information sharing while exposing institutions to risks such as system compromise, data breaches, and unauthorised access (Duque & Omar, 2015). In response, financial security emphasises confidentiality, integrity, and availability of stored data. Ensuring business continuity requires resilient, affordable security systems that reduce economic and organisational impact of cyberattacks (Von Solms & Van Niekerk, 2013). One framework, the System of Systems (SoS), integrates multiple security systems to identify evolving risks and implement proactive defence mechanisms.

Traditional cybersecurity measures within the banking sector include Device management via consistent updates; Organisational risk management; and Investment in employee training. Abdulrahman (2020) notes popular strategies by financial executives to lower cyber risks mandatory awareness training and allocating sufficient funds for security. Skills shortages among cybersecurity employees, particularly in technical expertise and communication, remain a significant challenge. The incorporation of AI into cybersecurity strategies reduces errors and expenses while enhancing security. Banks employ ML, deep learning (DL), and natural language processing (NLP) to detect and prevent cyber threats (Prince et al., 2024). These tools can process vast amounts of data and identify patterns, thereby enhancing the efficiency of cyber threat detection. AI techniques facilitate rapid detection and prevention of risks, ensuring high security standards to prevent data breaches. The incorporation of AI has enhanced archaic cybersecurity techniques in banking employing encryption software to keep cybercriminals away from their confidential information and online transactions (Mhlanga, 2020). Over time, ML algorithms adapt by learning from past data, enabling them to recognise anomalies and predict attacks more effectively than static, rule-based systems.

Banks prioritise cybersecurity due to sensitive client data. Safeguards include ATM validation, OTP confirmation, and pre-enrolment with signature authorisation. AI, while beneficial for cybersecurity, introduces adversarial attacks that manipulate training data to deceive ML models (Prince et al., 2024). Despite risks, AI is significant in detecting malware and ransomware early, automating monitoring, and freeing security professionals for strategic tasks. The reliance on AI in cybersecurity fuels rapid expansion of AI-driven security markets, reflecting demand for adaptive defence systems. AI is positioned not merely as a supplementary tool but central to modern banking cybersecurity.

2.2 Overview of Machine Learning Techniques for the Financial Sector

Machine learning techniques (MLTs), illustrated in figure 1, are critical in protecting banking systems. ML is used to predict loan approvals, detect fraudulent transactions, and monitor anomalies (Nuthalapati, 2023). Financial institutions deploy ML and DL against sophisticated cyber threats, mirroring technologies used by cybercriminals.

Modern attacks demand adaptive and proactive strategies (S. Mishra, 2023). ML enables real-time threat detection through analysis of large datasets, anomaly identification as early warnings, and continuous adaptation to emerging cyber threats (Prince et al., 2024). However, ML introduces vulnerabilities. Adversarial ML can distort threat assessments, hence creating vulnerabilities in security systems. To mitigate this, institutions train ML algorithms on accumulated data to identify behavioural patterns associated with breaches, strengthening predictive accuracy. AI automates detection and response, enhancing efficiency. Subcategories include: DL uses neural networks to detect anomalous network traffic patterns and identify malware. NLP enhances threat intelligence in the deployment of self-response systems, i.e. chatbots, and text retrieval. Reinforcement learning modifies defence tactics identifying new environmental interactions and improving security measures. Behaviour analytics use User and Entity Behavioural Analytics (UEBA) identify insider threats, unauthorised account compromises, or other suspicious behaviour.

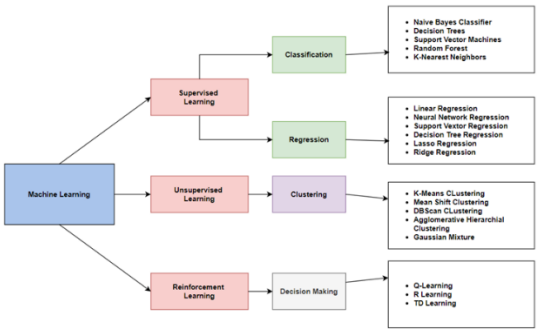


Figure 1: Machine Learning Techniques in Financial Sector

Supervised learning forecasts and categorises data, with Naïve Baye Classifier (NBC), Decision Tree (DT), Support Vector (SV), Linear Regression (LR), and K-Nearest Neighbours (KNN) as examples. Unsupervised learning techniques investigate hidden patterns in data without training datasets, labelled data, or predicted results. The main objective of this learning technique is for data clustering, often via Mean shift Clustering (MSC), and K-Means Clustering (KMC). Reinforcement learning strategies observe the system's environment, execute control, and optimise cumulative rewards. Q learning, R-Learning, and TD-Learning are instances of reinforcement learning algorithms.

ML algorithms analyse historical data to detect deviations from normal user behaviour, network traffic, or system activity patterns. ML enhances adaptability and self-learning of digital banking systems, strengthening security. Such adaptability ensures persistent protection of sensitive assets (S. Mishra, 2023). Digital banks benefit from incorporating ML into their cybersecurity protocols, enabling rapid threat response and safeguarding assets. ML-based cybersecurity solutions in digital banking should also address ethical issues that come with their use in the financial sector. MLTs in finance are applied in data collection, quality checks, cleansing, feature engineering, training and validation, monitoring, incident response, evaluation, updates, and privacy protections.

3. Method

In this section the methodology employed to conduct the study and obtain the findings is described. It provides an overview of the study's tools and research design. The choice of research approach is justified in this section. A real-world case study is incorporated to contextualize the experimental setup and validate the applicability of the proposed approach to practical financial environments. The implementation of network analysis using Wireshark and the creation of a secure cyber financial network using GNS3 are described. It presents practical work employing Kali-Linux to launch targeted network cyberattacks. Data is gathered from the network under analysis both prior to and during the attacks. The procedures used to develop the ML algorithm for identifying and preventing cyberattacks are thoroughly illustrated in this section.

3.1 Research Design

The technologies currently employed in the banking industry and cybersecurity procedures that are required are investigated in this study using both a qualitative and quantitative methodology.

Quantitative research is the methodical study of phenomena through the collection of quantifiable data and the use of statistical, mathematical, or computational methods (Recker, 2021). Researchers and statisticians employ mathematical frameworks and concepts related to the quantity under investigation in the research approach. This research will benefit from a quantitative approach, as its objective is to identify vulnerabilities and develop an algorithm for cybercrime defence within financial services systems at a network level. The study utilises Kali Linux to simulate real-world cyber threats and vulnerabilities in the banking industry. The study also utilises Kali Linux's penetration testing tools to collect data on attack vectors, analyse security gaps, and iteratively refine the algorithm based on observed patterns and insights. An actual financial network will be replicated by an emulated cyber security financial network. Graphical Network Simulator-3 (GNS3) will be used as a simulation tool or software for designing and creating the financial network. GNS3 is an open-source platform that is used to record data traffic in real time and to construct intricate networks. Wireshark, which establishes a path for the data MTL method, will be used to analyse network data flow.

Recker (2021) asserts that a qualitative approach is useful for contextualising real-life events. A qualitative methodology is applicable in this study, as its objective is to explore cybersecurity in the context of the financial services industry at a network level. This study will investigate a case study on what systems are in place to track, monitor, protect and respond to cyber threats. This approach enables the study to investigate the cybersecurity measures financial institutions have implemented on their systems at a network level, be it online or offline protocols, machines or people. The investigation aids in identifying the optimal cybersecurity measures necessary for the existing networks. The experimentation process involves simulating controlled scenarios within a standardised financial sector network to systematically collect data for ML applications. Python is utilised for data analysis, model development, and performance evaluation in ML applications. To facilitate these tasks, the open-source web-based tool Jupyter Notebook is employed, offering a comprehensive suite of libraries and frameworks specifically designed for such purposes.

This study focuses on enhancing cybersecurity within the financial sector through the application of ML algorithms. It is important to emphasise, however, that the aim is not to completely redesign or replace existing cybersecurity frameworks, nor to conduct comparative analyses. Furthermore, it should also be noted that this study did not involve actual physical penetration testing on live networks.

3.2 Case Study: Cyber-Attacks and Cybersecurity Readiness of Iraqi Private Banks

The increasing digitisation of financial services has brought unprecedented convenience to consumers while simultaneously exposing financial institutions to escalating cybersecurity threats. According to the World Bank's Global Findex Database, global account ownership rose from 51% in 2011 to 76% in 2021, with digital payments usage in emerging economies increasing from 35% to 57% over the same period (World Bank Group, 2022). Iraq, as part of this global trend, has witnessed a growing reliance on digital financial services, making its banking infrastructure a target for cybercriminals. Against this backdrop, (Hasan et al., 2021) investigated the cybersecurity readiness of Iraqi private banks, providing a representative case of the challenges faced by financial institutions in emerging economies.

The study adopted a questionnaire-based methodology to compensate for the limited availability of official data on cyber incidents within Iraq's banking sector. A total of 51 responses were collected from clients residing in Baghdad, Karbala, Babel, and Najaf, thereby reflecting perspectives from central regions of the country. The demographic profile of respondents indicated a male majority, 62.7%, with the largest age group being 36-45 years, 51%. Furthermore, more than half of participants reported four to six years of technology use, suggesting a reasonably experienced user base. The investigation focused on five primary categories of cyber threats: cyberstalking, hacking, phishing, cross-site scripting (XSS), and DDoS attacks. Cybersecurity readiness was conceptualised as the dependent variable, measured through eight questionnaire items, while each cyber threat represented an independent variable. Findings revealed that cyberstalking was a recurring issue, with consumers reporting experiences of harassment, blackmail, and extortion attempts while accessing banking services. Hacking was also perceived as a major risk, particularly in relation to the use of insecure applications and public Wi-Fi, though actual instances of credential theft were less frequent. Phishing emerged as one of the most significant concerns, with participants highlighting exposure to fraudulent links, deceptive emails, and counterfeit websites designed to harvest sensitive information. Similarly, XSS vulnerabilities were evident, with users frequently redirected to malicious websites through misleading content, while DDoS attacks reflected customer fears regarding the capacity of banks to ensure data protection and service continuity.

Statistical analysis, however, indicated no significant relationship between cybersecurity readiness and the independent variables, with hacking even demonstrating a negative correlation. This suggests that, despite growing awareness of cyber threats, Iraqi banks' defensive strategies have not instilled sufficient consumer confidence. Respondents consistently expressed scepticism regarding the sector's ability to secure online transactions, citing the prevalence of insecure public internet connections and the frequency of cyberattacks as key factors undermining trust.

The study's conclusions are prominent in the post-pandemic context, as financial pressures and increased digital reliance have amplified cybercrime globally. Iraqi banks, while actively working to improve their cybersecurity infrastructure, remain constrained by persistent vulnerabilities and a gap between perceived risks and actual preparedness. The case study underscores the pressing need for enhanced technical support, robust frameworks, and proactive strategies to address the evolving threat landscape. Through situating Iraq's experience within the broader discourse on cybersecurity in financial institutions, this case study highlights both the localised challenges of emerging economies and the universal need for adaptive, technology-driven solutions. Building on these findings, subsequent research proposes an AI-based cybersecurity framework

designed to strengthen detection and mitigation capabilities, thereby bridging the gap between institutional readiness and the sophistication of modern cyberattacks.

3.3 Research Instruments

This study employed a suite of simulation, virtualisation, penetration testing, and data analysis tools to design, evaluate, and secure a financial network environment. The primary setup comprised GNS3 and VMware, integrated with Kali Linux, supported by an 11th Gen Intel Core i7 processor, 8 GB RAM, Windows 11 (64-bit) operating system. GNS3 served as the main platform for emulating a financial network. It enabled the design and testing of complex topologies using virtual and physical devices. GNS3 is valued for its flexibility, scalability, and cost-effectiveness, supporting devices from Cisco, Juniper, MikroTik, and others. For this study, GNS3 version 3.0.5 was used in conjunction with VMware for improved performance. VMware Workstation Player (v17.6.3) was employed as the preferred virtualisation tool due to its efficiency and support for multiple operating systems. Its compatibility with GNS3 and free availability made it optimal for the study's requirements.

Kali Linux, an open-source Debian-based operating system, provided penetration testing capabilities to simulate cyberattacks. Its suite of tools, such as Yersinia, Ettercap, hping3, Nmap (network mapping), Netcat (port scanning), Scapy (packet manipulation), and Ping was used to exploit vulnerabilities in the emulated network. Wireshark, a packet analyser, captured and analysed traffic within the simulated network, exporting data for pre-processing and further analysis. For data processing and machine learning, Python was selected due to its readability and extensive libraries. Key modules included NumPy, SciPy, Pandas, Scikit-learn, Seaborn, and Logging, enabling statistical computation, data transformation, model development, and visualisation. Jupyter Notebook facilitated implementation, combining live code, analysis, and documentation. Collectively, these instruments ensured a robust experimental environment for simulating, analysing, and mitigating cybersecurity threats in financial systems at a network level.

3.4 Bank Architecture

Modern banking architecture is essential for cross-enterprise intelligence, analytics, compliance, and security (Keen et al., 2018). The rise of smartphones, blockchain, cloud computing, AI, Internet of Things (IoT), and open APIs has driven intelligent digital banking, enabling customer-centric, technology-driven services. The financial sector now extends beyond traditional lending to include investments, portfolio management, and lifestyle services. Emerging technologies such as AI-powered chatbots, robot-advisors, fraud detection, automation, and gamification enhance personalisation and risk management. Big Data supports smart banking, segmentation, and churn analysis, while IoT enables account automation and real-time services. Blockchain applications include open banking APIs, digital wallets, smart contracts, and central bank digital currencies. Biometrics and cloud computing further strengthen authentication and core banking flexibility.

The study adopts a Service-Oriented Architecture (SOA), depicted in figure 2, enabling interoperability and integration across distributed systems. SOA's modular, reusable, and open-standards-based design overcomes the rigidity of legacy ICT infrastructures, facilitating adaptability, seamless application integration, and strategic enterprise value (Meredith & Bjorg, 2003).

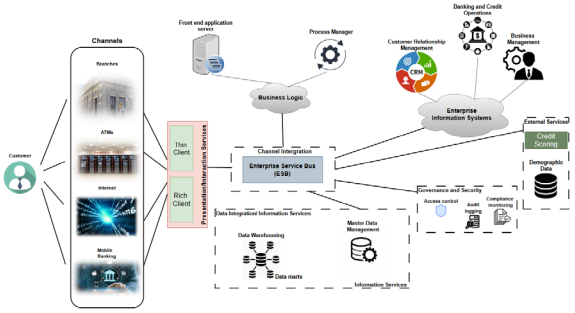


Figure 2: Bank Architecture

3.5 Cybersecurity of Bank Architectures and AI Integration

Cybercrime in banking encompasses identity theft, fraud, malware, and intrusions aimed at disrupting operations or stealing sensitive data (S. Mishra, 2023). Traditional security strategies often struggle with scalability and advanced threats, highlighting the need for comprehensive cyber-risk management. Banks rely on encryption, firewalls, intrusion detection systems (IDS), and mobile security to safeguard customer data and

maintain trust. A robust security program integrates policies, protocols, and tools to ensure confidentiality, integrity, and availability while complying with financial regulations. The components of the financial sector cybersecurity, include secure payments, data protection, and mobile safeguards. AI enhances these frameworks by automating repetitive tasks, providing real-time threat detection, and strengthening predictive capabilities. Applications include spam filtering, fraud detection, biometric authentication, and behavioural analytics.

3.6 Security Concerns in Financial Networks

Financial networks face increasing cyber threats targeting both institutional IT infrastructure and customers, often compromising transaction data or monitoring systems. Among the most prevalent are DDoS attacks, which overwhelm servers with malicious traffic, disrupting core banking, payment processing, and ATM services. Man-in-the-Middle (MitM) attacks intercept and alter communication channels, enabling attackers to reroute funds, inflate credit limits, or suppress compliance alerts, thereby undermining operational integrity. Cryptojacking exploits institutional computing resources for unauthorised cryptocurrency mining, degrading system performance, increasing costs, and impacting transaction and payment systems. Credential stuffing uses automated tools and stolen login data to infiltrate financial systems, facilitating fraud, data theft, and reputational damage. To mitigate these risks, institutions emphasise unique credentials, anomaly monitoring, and multi-factor authentication (MFA).

3.7 Machine Learning Process

The proposed methodology is illustrated in Figure 3. Three datasets were prepared. Each processed into thirteen features and divided into training (70%), validation (30%), and a separate testing set. The model was trained and refined through hyperparameter tuning, with its primary function to classify network traffic and detect cyberattacks, and its secondary role to enforce firewall rules by blocking suspicious packets while permitting legitimate ones.

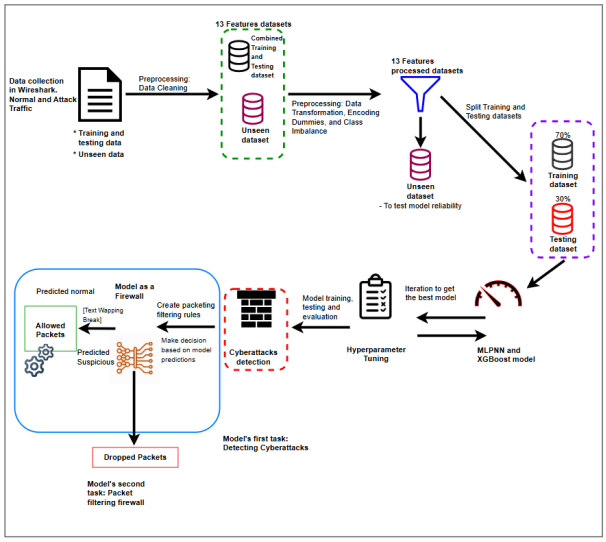


Figure 3: Model Architecture and Methodology

Data acquisition was central to the process. Real-time network traffic was recorded before and during simulated cyberattacks, including DDoS, MitM, False Data Injection Attacks (FDIA), and Credential stuffing. Wireshark was used to capture traffic, evident in figure 4, given its accuracy in packet analysis. Extracted features included IP and MAC addresses, ports, TTL values, protocol flags, and ICMP codes. Normal traffic was also recorded to establish baseline behaviour. Activities performed during acquisition included generating audit and transaction logs, data transfers, file uploads, file relocation, server access, and network reachability checks. This structured dataset enabled the model to learn patterns of normal versus malicious activity, ensuring robust evaluation across unseen scenarios.

Time	Src_IP	Dst_IP	Protocol	Src_Port	Dst_Port	Src_MAC	Dst_MAC	Time_to_Live	Flag	ICMP_Type	ICMP_Code	Packet_Cate
42292	10.10.10.100	10.10.10.15	HTTP	45378.0	8080.0	PCSSystemtec_b4a1:05	0cb1:3d:44:00:00	64	0x018	NaN	NaN	Mali
61375	10.10.10.15	10.10.10.100	HTTP	8080.0	45378.0	0cb1:3d:44:00:00	PCSSystemtec_b4a1:05	64	0x019	NaN	NaN	Mali
74663	10.10.10.100	10.10.10.15	HTTP	45388.0	8080.0	PCSSystemtec_b4a1:05	0cb1:3d:44:00:00	64	0x018	NaN	NaN	Mali
79185	10.10.10.15	10.10.10.100	HTTP	8080.0	45388.0	0cb1:3d:44:00:00	PCSSystemtec_b4a1:05	64	0x019	NaN	NaN	Mali
94515	10.10.10.100	10.10.10.15	HTTP	45392.0	8080.0	PCSSystemtec_b4a1:05	0cb1:3d:44:00:00	64	0x018	NaN	NaN	Mali
96714	10.10.10.15	10.10.10.100	HTTP	8080.0	45392.0	0cb1:3d:44:00:00	PCSSystemtec_b4a1:05	64	0x019	NaN	NaN	Mali
04574	10.10.10.100	10.10.10.15	HTTP	45404.0	8080.0	PCSSystemtec_b4a1:05	0cb1:3d:44:00:00	64	0x018	NaN	NaN	Mali

Figure 4: Sample of Dataset

3.7.1 Data preprocessing

Preprocessing was the foundational step, ensuring raw network traffic data was systematically organised, transformed, and cleansed for machine learning. This phase addressed issues such as noise, incompleteness, high dimensionality, and class imbalance to enhance model accuracy and reliability.

3.7.2 Data cleaning

Missing or incomplete values, particularly across protocol-specific fields, were handled without discarding rows. Instead, missing entries were substituted with zeros, retaining critical records while ensuring dataset consistency.

3.7.3 Data transformation

The dataset was split into training (70%) and validation (30%) subsets and transformed through feature engineering, normalisation, and encoding. IP and MAC addresses were hash-encoded, protocols one-hot encoded, and Boolean values converted to integers. A critical step was addressing the pronounced class imbalance in Packet_Category using SMOTE. Unlike random oversampling, SMOTE generated synthetic minority-class samples by interpolating between nearest neighbours, enhancing diversity and reducing overfitting risks. Stratified sampling, maintained class proportions across all subsets, ensuring robust and generalisable evaluation.

3.7.4 Model selection

The study evaluated four models: Decision Tree, Support Vector Machine (SVM), Multilayer Perceptron Neural Network (MLPNN), and Long Short-Term Memory (LSTM), to determine the most effective approach for detecting financial network cyberattacks. Given the dataset's complexity, the chosen model needed to handle heterogeneous variables, capture non-linear relationships, and deliver high interpretability for real-world cybersecurity.

The Decision Tree classifier emerged as the strongest performer, achieving 99.98% accuracy with precision, recall, and F1 scores all exceeding 99.97%. Its confusion matrix showed zero false positives and only 13 false negatives out of nearly 60,000 samples, minimising risks of undetected threats or false alarms. Beyond its accuracy, the model's interpretability was pivotal: top features such as Flag, IP/MAC addresses, ports, and TTL values provided transparent insights into malicious traffic patterns. The tree's structure illustrated clear, low-impurity decision paths, reinforcing analyst trust and supporting timely, accountable threat response. SVM, MLPNN, and LSTM also delivered strong performance, with accuracies ranging from 98.99% to 99.92%. However, SVM produced higher false positives, MLPNN slightly increased misclassifications, and LSTM, though accurate, lacked the Decision Tree's interpretability. Performance evaluation confirmed the Decision Tree's reliability, generalisation, and operational suitability. Its combination of accuracy, zero false positives, and transparent decision-making established it as the optimal model for financial cybersecurity at a network level.

4. Findings and Analysis

This section presents the findings derived from the previous sections. The section contains documentation of the analysis of the findings. The performance evaluation of the proposed machine learning algorithm in detecting and mitigating cyberattacks within financial systems at a network level and visual representations. It reviews datasets, pre-processing methods, literature insights, and model performance.

The literature revealed that the financial sector faces distinct and complex cybersecurity challenges at a network level, as it manages sensitive data and high-value transactions. Traditional, rule-based measures are increasingly inadequate against evolving threats, while AI and ML have become essential for real-time detection and response. However, challenges such as adversarial AI, a shortage of skilled personnel, and reliance on reactive strategies persist. Simulated environments like GNS3 offer controlled conditions for training AI models, underscoring the importance of predictive and proactive approaches.

The case study of private banks in Iraq highlighted weak institutional readiness despite growing awareness of cyber risks. Frequent threats, including phishing, DDoS, and hacking, were reported, yet responses remained ad hoc, lacking structured strategies. Respondents also identified unsecured networks, poor data confidentiality, and insufficient user training as major concerns, revealing a disconnect between policy and implementation.

The datasets were generated in a simulated financial network, illustrated in figure 5, using Wireshark, capturing attributes such as IP and MAC addresses, ports, flags, protocols, and ICMP codes. These were divided into

training, validation, and testing subsets to ensure reliable performance evaluation. Data pre-processing included cleaning and encoding, while SMOTE was applied to address class imbalance. By generating synthetic samples, SMOTE reduced overfitting and enhanced generalisation, contributing significantly to model robustness.

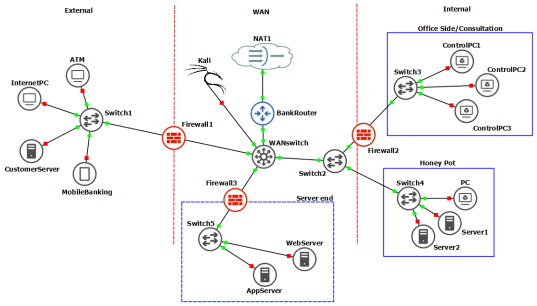


Figure 5: Hypothetical Bank network in GNS3

Among the tested algorithms, the Decision Tree proved most effective, achieving 99.98% accuracy with zero false positives and only 13 false negatives. Compared with SVM, MLPNN, and LSTM, it offered superior interpretability, efficiency, and compliance readiness. Feature importance analysis identified the Flag attribute as most critical, while the confusion matrix confirmed strong precision, recall, and balanced classification. These results establish the Decision Tree as the most suitable model for deployment in financial cybersecurity systems at a network level.

5. Conclusion

This study, supported by a real-world case of cyberattacks on Iraqi banks, developed a supervised machine learning model for proactive intrusion detection and prevention in financial networks. Using GNS3-simulated environments, Wireshark-captured traffic, and SMOTE pre-processed datasets, a Decision Tree algorithm achieved exceptional accuracy (99.98%), precision, and recall, outperforming SVM, MLPNN, and LSTM. The model's transparency ensures regulatory compliance while minimising false alarms. Key findings emphasise the need for proactive detection, AI integration, and resilience against threats such as phishing, ransomware, and DDoS. The research demonstrates AI's potential for real-time cyber defence, bridging theoretical innovation with practical financial security applications.

6. Recommendations

The proposed model should be tested in real-world financial networks through collaboration with industry stakeholders. Integrating real-time data analysis and continuous monitoring would enhance threat detection and response. Adaptive learning strategies should allow the model to update based on new data, addressing evolving cyber threats. Future work includes developing autonomous, real-time models acting as intelligent firewalls, incorporating reinforcement or online learning, and expanding simulations to address stealthy attacks and larger network topologies. Assessing performance under varying loads and applying explainable AI may help improve robustness, scalability, and transparency, supporting regulatory compliance and trust in automated financial cybersecurity systems.

AI Declaration: This study was independently prepared, authored, and compiled in its entirety without the use of any automated systems or artificial intelligence technologies.

Ethics declaration: The utilisation of the virtual simulation of a financial system was driven by its non-invasive nature, ensuring the safety of existing systems, human users, and security protocols. The study adhered to publicly available data and research for analysis and reporting, safeguarding the confidentiality of financial system security. This cautious approach ensured that the current cybersecurity operations of the financial system were not compromised. Consequently, all research methods and data incorporated in this study were conducted with the utmost objective and careful consideration of the subject matter.

References

Abdulrahman, S. (2020). The Significance of Cybersecurity System in Helping Managing Risk in Banking and Financial Sector. *Journal of Xidian University*, *14*(7). https://doi.org/10.37896/jxu14.7/174

Duque, S., & Omar, M. N. Bin. (2015). Using Data Mining Algorithms for Developing a Model for Intrusion Detection System (IDS). *Procedia Computer Science*, *61*, 46–51. https://doi.org/10.1016/j.procs.2015.09.145

Hasan, M. F., Alramadan, N. S., Al-Ramadan, N. S., & Professor, A. (2021). Cyber-attacks and Cyber Security Readiness: Iraqi Private Banks Case. *Social Science and Humanities Journal*, *05*. https://doi.org/10.6084/m9.figshare.15190185.v2

Keen, M., Kaushik, R., Singh Bhogal, K., Aghara, A., Simmons, S., Dulaney, R., Dube, S., & Allison, A. (2018). *Case Study: SOA Banking Business Pattern*.

Meredith, L. G., & Bjorg, S. (2003). Contracts and types. *Communications of the ACM*, *46*(10), 41–47.

Mhlanga, D. (2020). Industry 4.0 in finance: the impact of artificial intelligence (ai) on digital financial inclusion. *International Journal of Financial Studies*, *8*(3), 1–14. https://doi.org/10.3390/ijfs8030045

Mishra, S. (2023). Exploring the Impact of AI-Based Cyber Security Financial Sector Management. *Applied Sciences*, *13*(10), 5875. https://doi.org/10.3390/app13105875

Nuthalapati, S. babu. (2023). AI-Enhanced Detection and Mitigation of Cybersecurity Threats in Digital Banking. *Educational Administration: Theory and Practice*, 357–368. https://doi.org/10.53555/kuey.v29i1.6908

Prince, N. U., Faheem, M. A., Khan, O. U., Hossain, K., Uddin Prince, N., Alkhayyat, A., Hamdache, A., & Elmouki, I. (2024). *AI-Powered Data-Driven Cybersecurity Techniques: Boosting Threat Identification and Reaction Nanotechnology Perceptions AI-Powered Data-Driven Cybersecurity Techniques: Boosting Threat Identification and Reaction*. https://doi.org/10.13140/RG.2.2.22975.52644

Recker, J. (2021). *Scientific Research in Information Systems*. Springer International Publishing. https://doi.org/10.1007/978-3-030-85436-2

neu Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *Computers and Security*, *38*, 97–102. https://doi.org/10.1016/j.cose.2013.04.004

World Bank Group. (2022). *The Little Data Book on Financial Inclusion*. www.worldbank.org

Analysis of Cybersecurity Strategies Across Continents

Pekka Pirinen, Piia Perälä and Martti Lehto
University of Jyväskylä, Finland

pekka.j.pirinen@jyu.fi
piia.m.h.perala@jyu.fi
martti.j.lehto@jyu.fi

Abstract: Cyberthreats are a global challenge and because of that countries across the world have been establishing and developing national cybersecurity strategies (NCSS). NCSSs are a way in which countries can present their political and social approach to cybersecurity in the sense of national security. The main principles of a NCSS are addressing the current cybersecurity landscape that the country in question has, the goals of future national cybersecurity development and how to achieve that desired state of national cybersecurity affairs. Being part of the country's national, political and cultural landscape, the actual cybersecurity strategies naming policies may differ from country to country. Although the most common goals are similar between different states' cybersecurity strategies, they have different emphasis regarding the role of citizens, the private sector and the public sector in national cybersecurity. This study aims to analyze and compare NCSSs from different continents and recognize the focus areas in the three crucial target groups: citizens, private sector and public sector. NCSSs of 20 different countries from Africa, Asia, Europe, North America, Australia and Oceania and South America were analyzed. All the strategies were collected from public sources and represent the latest version of selected countries' cybersecurity strategies. The strategies were analyzed using the Commitment to Development (C2D) approach. The Democracy Index was utilized in the study to recognize and differentiate countries based on their index rating. The main findings of this study were that with a few exceptions countries sharing similar Democracy Index rankings have similar attitudes towards their citizens, private sector and public sector. The results also varied in a continental perspective, and some continents align better with each other than others.

Keywords: Cybersecurity strategy, Comparative analysis, Commitment to development, Citizens, Private sector, Public sector

1. Introduction

Cybersecurity in the national context has become a subject of interest to different countries from all around the world. Countries have started to build their own approach to national cybersecurity. These different national cybersecurity strategies (NCSSs) form a possibility to investigate the countries' strategic national ways of cybersecurity thinking. NCSSs are a part of the countries' national, political and cultural landscape and have different emphasis from country to country.

Previous research on NCSSs has approached the topic from various perspectives. For example, Kleiner (2025) focused on how different political systems influence national cybersecurity. While Carr (2016) analyzed the role of cooperation between the public and private sectors in NCSSs. In addition, NCSSs have been studied to gain understanding for the further development of strategies (e.g., Odebade & Benkhelifa, 2023; Shafqat & Masood, 2016).

Previous research has often focused on a single issue or perspective, such as governmental aspects (e.g., Odebade & Benkhelifa, 2023) or has been limited to specific geographic regions (e.g., Lehto, 2013). This study aims to provide a broader understanding of how countries emphasize different focus areas in their NCSSs. It differs from earlier work by examining how public-private partnership collaboration mechanisms are reflected in NCSSs and how the dimensions of the C2D approach are represented. Furthermore, the study includes strategies from countries across all continents, rather than being limited to one or a few regions.

In this study NCSSs from all the world's continents are analyzed and compared. The strategies are then evaluated by the three focus areas: citizens private sector and public sector. These three focus areas were then divided into different sub focus areas to enhance the analysis.

2. Public-Private Partnership Collaboration Mechanisms

Public-private partnership in cybersecurity can be seen as different ends and means in which governmental public sector collaborates with private sector thus enhancing the different aspects of cybersecurity in the country. The most crucial part of public-private partnership is the protection of cyber-related critical infrastructure, that is operated in one way or another by the private sector in every country of the world. Public-private partnerships collaboration mechanisms have their roots in the 1990's when countries started to privatize their national critical infrastructure for economic development (Carr, 2016).

Public-private partnerships collaboration mechanisms are visible in this study's NCSSs, and they range from economical and educational collaboration to protecting of critical infrastructure and national armed defense. In all the NCSSs some form of public-private partnerships could be observed that would concern the critical infrastructure or national economy.

3. The Commitment to Development (C2D) Approach

Commitment to Development (C2D) is an approach that focuses on cybersecurity from societal and economical point of view by combining different values into cybersecurity policies. The C2D enables countries to identify and act on different national needs and co-operation possibilities and works as a tool for understanding the cybersecurity field and its different aspects in national and international context (Kerttunen & Tikk, 2019).

The C2D consists of three different elements which are direction, strategic capacity and policy analysis. The direction element of C2D focuses on the societal values and political values behind the NCSSs and how the strategies represent them. Strategic capacity is a set of different aspects in the NCSSs that enhance the national capabilities offering sustainable security to the country. Policy analysis focuses on how the different NCSSs policies are defined into measurable metrics. (Kerttunen & Tikk, 2019.)

In the previous research, the C2D approach has been utilized by Busetti and Scanni (2025) where the C2D was used to analyze the national security role of securing critical infrastructure in cyberspace.

4. Methodology

The study material consists of NCSSs from 20 different countries across every continent. The strategies were selected in a way that at least three countries were included from each continent. The selection of countries was based on the Democracy Index, ensuring that within each continent, the chosen countries represented varying levels of democracy. The Democracy Index ranks countries in the scale of 1 being an authoritarian country to 10 being a full democracy. By utilizing the Democracy Index in the study, it is possible to investigate how the countries in different democracy levels view cybersecurity and how the different areas (citizens, private sector & public sector) contribute to cybersecurity. Table 1 presents the countries included in the study along with their Democracy Index rankings and the publication years of their NCSSs and the publisher of the strategy.

Table 1: Continents, countries, their 2024 Democracy Index rating and the publisher of the strategy.

Continent	Country	Democracy Index 2024	Year of Strategy Publication	Publisher
Asia	Saudi Arabia	2,08	2020	Kingdom of Saudi Arabia
Africa	Eswatini	2,6	2022	Eswatini Communications Commission
Asia	Pakistan	2,84	2021	Government of Pakistan
Africa	Tanzania	5,2	2022	President´s Office & Public Service Management and Good Governance
North America	Mexico	5,32	2017	Government of Mexico
South America	Paraguay	5,92	2025	Government of Paraguay
Australia & Oceania	Papua New Guinea	5,97	2024	The Independent State of Papua New Guinea
Europe	Romania	5,99	2022	Government of Romania
South America	Argentina	6,51	2024	Republic of Argentina National Executive Branch

Continent	Country	Democracy Index 2024	Year of Strategy Publication	Publisher
Asia	India	7,29	2020	Data Security Council of India
Africa	Botswana	7,63		Ministry of Transport and Communications
Europe	Latvia	7,66	2023	Ministry of Defense
Asia	Israel	7,8	2025	Israel National Cyber Directorate
North America	USA	7,85	2023	The White House
Asia	Japan	8,48	2021	The Government of Japan
South America	Uruguay	8,67	2024	The President of Uruguay
North America	Canada	8,69	2025	Ministers of Public Safety and Emergency Preparedness
Australia & Oceania	Australia	8,85	2023	Australian Government
Europe	Finland	9,3	2024	Prime Minister´s Office
Australia & Oceania	New Zealand	9,61	2019	New Zealand Government

Content analysis by Weber (1990) was used to identify the content of the strategies. Atlas.ti software was used for the analysis. The strategies were uploaded to Atlas.ti and coding was made based on the three different focus areas and their sub areas. For the analysis, the C2D approach was utilized. The analysis was conducted in three phases.

In the first phase, the strategies were examined to identify how the public-private partnership collaboration mechanism is manifested at the strategic level. The analysis focused on how citizens, the private sector, and the public sector are represented in NCSSs. For example, the roles that these different actors play in national cybersecurity.

In the second phase, the public-private partnership was analyzed in the context of the countries' Democracy Index rating. After that the NCSSs were compared within and between continents.

In the third phase, the public-private partnership was analyzed in more depth to understand how the dimensions of the C2D approach and most specifically its direction aspect was reflected in this collaboration mechanism.

5. Research Results

5.1 Citizens

Four sub focus areas were utilized to differentiate the countries' emphasis on citizens between each other. The sub focus areas are how the NCSSs consider *the citizens in legislation* (legislation affecting citizens), *the individual citizens' rights* (mentions about citizens' rights), *citizen agency* (mentions of citizens agency) and *the safety and privacy of citizens* (mentions about citizens safety or privacy).

Asian countries varied from each other. Pakistan does not view its citizens as an active agency and emphasizes the importance of securing the cyberspace of its citizens (Government of Pakistan, 2021). India has also a similar security first perspective to its citizens in the cyberspace. The focus of India´s strategy considering its citizens is to secure the country´s cyber infrastructure for its citizens in the normal and emergency conditions (Data Security Council of India, 2020).

Saudi Arabia has a business-oriented attitude on its citizens. All the citizens' agencies are based on their business attributes (Kingdom of Saudi Arabia, 2020). Israel emphasizes that the citizens of Israel are active participants in

the cyberspace and regarding the national security of the country every citizen has a duty to protect the nation in emergency situations. Israel also brings up the role of democracy regarding its citizens. (Israel National Cyber Directorate, 2025.) Japan emphasizes that its citizens as active agency should have the opportunity to participate in a democratic and rights-respecting protected cyberspace. Japan sees its citizens as a crucial factor in building the country´s secure cyberspace. (The Government of Japan, 2021.)

The European countries Finland, Latvia and Romania have very similar relation to their citizens role in their NCSSs. Latvia sees its citizens as agency that have clear roles in national cybersecurity in the case of national emergency. Latvia also emphasizes that citizens' rights should be protected in all circumstances. (Ministry of Defense, 2023.) Finland also sees its citizens as a crucial part of the whole society as a cybersecurity provider and emphasizes the citizens role and responsibility in the civilian defense of the national cyberspace. Finland also emphasizes that the government should ultimately guarantee the citizens' rights and security in the cyberspace. (Prime Minister´s Office, 2024.) Romania sees its citizens as active agency and emphasizes that every Romanian citizen has a role in the national cyber defense. Romania also emphasizes that every citizen should be responsible for their actions in the cyberspace. (Government of Romania, 2022.)

The countries in the Australia and Oceania have similarities and differences in their emphasis on their citizens roles in cybersecurity. Australia views its citizens as agency but emphasizes that the government should have a stronger role in protecting their citizens and their rights in the cyberspace (Australian Government, 2023). Papua New Guinea emphasizes that the rights and security of its citizens in the cyberspace should be protected, and citizens need to assist the government in securing the cyberspace. Papua New Guinea also emphasizes that the level of its citizens cybersecurity skills should be improved to improve the country´s national cybersecurity situation. (The Independent State of Papua New Guinea, 2024.) New Zealand emphasizes that the individual citizen is an agency and a cybersecurity provider to the country. New Zealand also emphasizes the citizens' rights and states that the country needs the contribution of its citizens to secure the national cyberspace. (New Zealand Government, 2019.)

The South American countries have an emphasis on focusing to protect the rights of their citizens in the cyberspace. Argentina emphasizes that the citizens privacy and rights need to be secured but does not view an individual citizen as a major agency and emphasizes the government´s role in building secure national cyberspace (Republic of Argentina National Executive Branch, 2024). Uruguay emphasizes that citizens are a crucial part of building the strategy and one of the objectives is to secure the citizens' rights and privacy in the national cyberspace. Uruguay also emphasizes that the citizens should also have responsibility for their own actions in the cyberspace. (The President of Uruguay, 2024.) Paraguay emphasizes need of citizens participation in the planning of the countries cyber policies. Paraguay also acknowledges that the country´s current legislation is not efficient enough to protect the citizens' rights and privacy in the cyberspace. (Government of Paraguay, 2025.)

In Northern American countries, there were major similarities between USA and Canada. Mexico has differences in its focus on its citizens. USA emphasizes that its citizens have too great responsibility for protecting the national cybersecurity and that the government should do more to protect the citizens in cyberspace. USA also highlights the role of co-operation between the citizens, and government to tackle cybersecurity issues. (The White House, 2023.) Canada also emphasizes that citizens have a too great role and responsibility in national cybersecurity. Canada sees cybersecurity as a whole society project and citizens as agency can do their part in protecting their own actions in the cyberspace. (Ministers of Public Safety and Emergency Preparedness, 2025.) Mexico emphasizes in its strategy that the government should protect the citizens' rights and privacy in the cyberspace and that citizens were involved in the making of the strategy. Mexico also acknowledges that the cyber maturity level of its citizens is not yet as high as it should be. (Government of Mexico, 2017.)

The countries in Africa have also similarities and differences in their attitudes towards their citizens. Botswana sees that the national cybersecurity is a whole society project, and its citizens are an agency and that the government's responsibility is to ensure an open and protected cyberspace for its citizens (Ministry of Transport and Communications, 2018). Tanzania recognizes that the country´s and its citizens cyber maturity level is low and that the citizens of the country are in a need of trustworthy cybersecurity infrastructure. The citizens of Tanzania are seen as a beneficiary of the strategy. (President´s Office & Public Service Management and Good Governance, 2022.) Eswatini also recognizes the low level of the country's citizens cyber maturity but also emphasizes that the citizens are responsible for producing national cybersecurity on their part (Eswatini Communications Commission, 2022).

5.2 Public Sector in the Role of National Defense and Armed Forces

The role of public sector regarding the countries' cyber defense was also one of the sub focus areas that were evaluated in the NCSSs. The role of public sector in cyber defense area differed between countries.

Asian countries have different attitudes regarding public sector and its role in national cyber defense. India and Saudi Arabia emphasize means of public civil defense in their NCSSs (Kingdom of Saudi Arabia, 2020; Data Security Council of India, 2020). Pakistan emphasizes in its strategy that the country´s Mil-CERT-teams should contribute to the security of public sector (Government of Pakistan, 2021). The two countries that stand out in Asia are Japan and Israel. Japan emphasizes that a cyberattack against Japan could be compared to an armed attack and the armed forces is ready to defend Japan (The Government of Japan, 2021). Israel emphasizes that the whole public sector works as a part of national defense with the armed forces and the country is willing to proactively act against cyber domain adversaries (Israel National Cyber Directorate, 2025).

European countries view the role of public sector and armed forces similarly. Romania emphasizes that the country´s national cyber defense should be based on NATO standards (Government of Romania, 2022). Finland and Latvia have more comprehensive approach. Finland emphasis that cyber defense is way to secure the operating conditions of the armed forces and recognizes that the whole society´s cyber defense can´t be under the responsibility of the armed forces (Prime Minister´s Office, 2024). Latvia emphasizes that in the case of emergency the armed forces would protect the society (Ministry of Defense, 2023).

Countries in Australia and Oceania also had different views regarding the role of public sector in the national cyber defense. New Zealand does not have a defensive perspective in its strategy. Australia only emphasizes the civilian defense role of its public sector (Australian Government, 2023). Papua New Guinea has the only military dimension regarding its public sector describing that the country´s Defence Cyber Intelligence Unit protects the country´s sovereignty in cyberspace (The Independent State of Papua New Guinea, 2024).

South America has a very similar way of viewing the role of public sector in the cyber defense area. In Argentina the role of the country´s armed forces is to protect country´s national cyberspace (Republic of Argentina National Executive Branch, 2024). Uruguay and Paraguay both emphasize that their armed forces protect their national cyberspace but also need to upgrade their cyberoperation capabilities to protect their sovereignty (The President of Uruguay, 2024; Government of Paraguay, 2025).

Canada and USA in North America have a similar way of viewing the role of public sector providing cybersecurity in the national level. USA emphasizes that the DoD if needed can pre-emptily disturb and neutralize cyberthreats (The White House, 2023). Canada also emphasizes that the armed forces jointly engages in operations that ensure the country´s safety in cyberspace (Ministers of Public Safety and Emergency Preparedness, 2025). Mexico sees the role of public sector in cyber defense as protecting the nations critical infrastructure (Government of Mexico, 2017).

African countries also differ in their attitudes towards public sector role in cyber defense. Tanzania does not have a cyber defensive perspective. Botswana emphasizes that the MoD is responsible for the country´s national cybersecurity (Ministry of Transport and Communications, 2018). Eswatini emphasizes that the country´s armed forces protect the country´s national cyberspace by cyberwarfare is necessary (Eswatini Communications Commission, 2022).

5.3 Private Sector

Five different sub focus areas were utilized to differentiate the countries between each other. The sub focus areas are *private sector´s role in national defense*, *private sector under state guidance*, *private sector in co-operation with the state*, *private sector´s security* and *private sector´s liability*.

Countries in Asia have differences and similarities towards the role of private sector. Pakistan emphasizes that the state should guide the private sector in favor of the national and the private sector´s security (Government of Pakistan, 2021). India and Saudi Arabia have a business-oriented perspective to private sector, but they also emphasize on strong state lead (Kingdom of Saudi Arabia, 2020; Data Security Council of India, 2020). Japan and Israel view that private sector and the state need to work in co-operation to protect the national security and economy (The Government of Japan, 2021; Israel National Cyber Directorate, 2025).

The European countries Finland, Latvia, and Romania have a similar perspective on private sector. Finland emphasizes the voluntary co-operation to improve the country's overall cybersecurity and national cyber defense (Prime Minister´s Office, 2024). Romania views the private sectors roles similarly as Finland but also

emphasizes the liability of private sector in national security (Government of Romania, 2022). Latvia also views the roles of private sector similarly emphasizing the private sector's role and liability in the national defense sector (Ministry of Defense, 2023).

Countries in the Australia and Oceania have also differences towards the role of private sector in national cybersecurity. Australia and New Zealand view that the private sectors focus is on the cybersecurity markets and in strong voluntary co-operation (Australian Government, 2023; New Zealand Government, 2019). Papua New Guinea views the role of the private sector being under the state´s guidance in building and developing the country´s cyber capabilities (The Independent State of Papua New Guinea, 2024).

The South American countries Argentina, Paraguay & Uruguay have a very similar attitude towards the role of private sector in national cybersecurity landscape. All the countries emphasize the voluntary co-operation aspect focusing on the economic aspects of this co-operation (Republic of Argentina National Executive Branch, 2024; The President of Uruguay, 2024; Government of Paraguay, 2025).

In North America there are similarities between USA and Canada as Mexico differs in its view of private sectors role in national cybersecurity. USA emphasizes co-operation, and the private sector should have liabilities on national cybersecurity (The White House, 2023). Canada also emphasizes on co-operation but also points out that the state can guide the private sector if necessary (Ministers of Public Safety and Emergency Preparedness, 2025). Mexico emphasizes that the private sector needs the guidance of the state to improve their cyber capabilities (Government of Mexico, 2017).

The African countries have also different viewpoints into the role of private sector in national cybersecurity. Tanzania emphasizes that the country´s cyberspace resorts to private sector and Botswana emphasizes the role of voluntary co-operation in building the country´s cyber capabilities (President´s Office & Public Service Management and Good Governance, 2022; Ministry of Transport and Communications, 2018). Eswatini emphasizes that private sector needs to be guided by the state to accomplish a secure private sector and national cyberspace (Eswatini Communications Commission, 2022).

6. C2D Capacity and the Different Sectors

While evaluating the role of three focus areas through the C2D approach and specifically its direction section, it can be indicated that the countries and their strategies follow the framing that C2D approach sets. According to the direction section of C2D it´s essential to countries that their cybersecurity thinking should be based on their nations' fundamental base values and national development policies (Kerttunen & Tikk, 2019). The countries emphasized their own national fundamental values which ranged from economic growth and individual rights to national defense and democracy. Every country differs from each other, and the evaluated strategies had strong indications of the country's direction regarding their national cybersecurity.

7. Key Themes and Research Findings

When reviewing the roles that countries give to their citizens there are clear differences and similarities between the countries based on their DI ranking. Highly ranked countries like New Zealand and Finland emphasize that their citizens are an active agency and countries with a lower DI ranking like Pakistan and Saudi Arabia emphasize that their citizens are not as strong agency and more control and protection from the state is needed.

If we look at the countries emphasis on their citizens' role from continental viewpoint, there are similarly aligned and differing continents. European countries emphasize their citizens´ role in building and securing their national cyberspace. South American countries emphasize on their citizens' rights and privacy in cyberspace. In North America, Canada and the USA have a similar attitude towards their citizens where as Mexico´s attitude towards its citizens is similar the South American countries. African countries common factor is that citizens are a developing and increasingly stronger part of the national cybersecurity. Australian and Oceanian countries emphasize that, if necessary, the state will protect the citizens in cyberspace respecting their rights. Asian Israel and Japan have a similar emphasis on their citizens viewing them as strong agency while promoting the role of rights and democracy. Other Asian countries India, Pakistan, and Saudi Arabia focus on their citizens role as being not as a strong agency focusing the citizens security and role in the national economy.

When reviewing the role of public sector in national cyber defense and armed forces, the DI ranking does not reflect into the countries' emphasis on the matter. Highly ranking countries like Finland and Australia do not share a similar view of public sector in armed cyber defense and lower ranking countries like Pakistan and Eswatini have similarities in their view of public sectors role in armed cyber defense.

The continents can be divided into similarly aligning and differentiating. South American countries similarly view the role of public sector in the cyber defense and their armed forces in the cyberspace. The European countries also have a similar basis on their public sectors' role in the cyber defense area. In North America, the USA and Canada have similar emphasis on the public sectors role as Mexico differs from the two. In Australia and Oceania, Australia and New Zealand do not emphasize the armed side of public sector whereas Papua New Guinea acknowledges it. Countries in Africa also fall into two categories where Tanzania does not recognize the armed side and Eswatini and Botswana have mentioned the armed side of public sector in cyberspace. The Asian countries have the most variance. India and Saudi Arabia do not discuss the armed side of public sector. Pakistan, Israel, and Japan all discuss the public sector's role in armed cyber defense.

The role of private sector does align with the DI rankings with some exceptions. Countries with a higher DI ranking like Australia and the USA emphasize the meaning of co-operation with the private sector in cybersecurity whereas lower ranking countries like Eswatini and Saudi Arabia emphasize the role of state lead private sector in cybersecurity.

The continents share similar differences and similarities in the role of private sector as previously have been discussed. In Africa Botswana and Tanzania emphasize more on voluntary co-operation whereas Eswatini emphasized the role of state lead and security. In Australia and Oceania, the countries focus on security side of private sector and Australia and New Zealand emphasize the voluntary co-operation whereas Papua New Guinea focuses on more developing the country´s private sector under state guidance. The European countries have a very similar perspective on private sector and its role in national cybersecurity emphasizing the role of voluntary co-operation. The South American countries also emphasize the role of voluntary co-operation. In North America Mexico differs from the USA and Canada by emphasizing the role of state lead in developing the country´s cyber capabilities, whereas the USA and Canada focus more on voluntary co-operation.

When observing the NCSSs through the C2D approach and its direction section one can argue that the different countries' base national values and policies are visible in all the three observed focus areas. Countries that resemble each other in their political and ideological values emphasized similar attributes regarding the focus areas. For example, when looking through the continental viewpoint all the European countries have similar attitudes in all the three focus areas.

When looking at the countries through their DI rating it can be argued that the countries sharing similar DI ranking also have similar attitudes towards the three focus areas. Highly functioning democracies share similar attitudes and countries with a more authoritative system have similar attitudes towards focus areas. All of this is based on their direction, and it shows clearly.

8. Implications and Future Research

Alignment between the countries' DI ranking and their relation to their national cybersecurity can be observed in two of the three focus areas. In the role of citizens and private sector there can be seen a connection between the DI ranking and the strategies C2D direction. However, there could be other reasons for the connection that this study does not investigate as thoroughly like the European countries having alliances in the EU and NATO.

When comparing this study to the previous research, similar and differing results can be seen. What differs this study is the utilization of Democracy Index ranking and the C2D approaches direction aspect and the combination of civil national cybersecurity into the armed side of national cybersecurity.

Possible future research toward this style of cybersecurity study could have a larger proportion of countries and possibly also take advantage of the older versions of their NCSSs creating a continuum to more strongly understand the direction behind their different focus areas.

AI declaration: No AI was used in this study.

Ethics declaration: No ethical clearance was needed for this study.

References

Australian Government. (2023) *2023–2030 Australian Cyber Security Strategy*. https://www.homeaffairs.gov.au/cyber-security-subsite/files/2023-cyber-security-strategy.pdf

Busetti, S. & Scanni, F.M. (2025*) Evaluating cybersecurity strategies in the energy sector. Evidence from Italy.* Rivista di Digital Politics, 1, pp.165–194. https://doi.org/10.53227/117428

Carr, M. (2016) '*Public–private partnerships in national cyber-security strategies*', International Affairs, 92(1), pp.43–62. doi:10.1111/1468-2346.12504.

Data Security Council of India. (2020) *National Cyber Security Strategy 2020.* https://www.dsci.in/files/content/knowledge-centre/2023/National-Cyber-Security-Strategy-2020-DSCI-submission.pdf
Eswatini Communications Commission. (2022) *Eswatini National Cybersecurity Strategy 2022-2027* https://www.esccom.org.sz/publications/reports/docs/Eswatini_National_Cybersecurity_Strategy_2022-2027.pdf
Government of Mexico. (2017) *National Cybersecurity Strategy*. https://www.gob.mx/cms/uploads/attachment/file/399655/ENCS.ENG.final.pdf
Government of Pakistan. (2021) *National Cyber Security Policy 2021*. https://moitt.gov.pk/SiteImage/Misc/files/National%20Cyber%20Security%20Policy%202021%20Final.pdf
Government of Paraguay. (2025) *Estrategia Nacional de Ciberseguridad 2025–2028*. https://mitic.gov.py/eoj0cad9uplo/2025/05/ENC-Paraguay-2025-2028-Mayo-20251558.pdf
Government of Romania. (2022) *Strategiei de securitate cibernetică a României, pentru perioada 2022–2027*. https://securitatea-cibernetica.ro/documente/Strategia-de-securitate-cibernetica-a-Romaniei.pdf
Israel National Cyber Directorate. (2025) *Israel National Cyber Security Strategy 2025*. https://www.gov.il/BlobFolder/news/cyber_strategy_2025/en/israel_national_cybersecurity_strategy_feb2025.pdf
Kerttunen, M. & Tikk, E. (2019) *National Cyber Security Strategies: Commitment to Development,* Cyber Policy Institute, 1, pp. 1–14.
Kingdom of Saudi Arabia. (2020) *National Cybersecurity Strategy*. https://nca.gov.sa/national_cybersecurity_strategy-en.pdf
Kleiner, J. (2025) '*How political regimes affect national cybersecurity: the polity flux effect'*, Democratization, 32(5), pp. 1181–1212. doi: 10.1080/13510347.2025.2451951.
Lehto, M. (2013) *The Ways, Means and Ends in Cyber Security Strategies.* In: Proceedings of the 12th European Conference on Information Warfare and Security, pp.182–190.
Ministers of Public Safety and Emergency Preparedness. (2025) *Canada's National Cyber Security Strategy: Securing Canada's Digital Future*. https://www.publicsafety.gc.ca/cnt/rsrcs/pblctns/ntnl-cbr-scrt-strtg-2025/ntnl-cbr-scrt-strtg-2025-en.pdf
Ministry of Defense. (2023) *The Cybersecurity Strategy of Latvia 2023–2026*. https://www.mod.gov.lv/sites/mod/files/document/Kiberdrosibas%20strategija%202023%20ENG.pdf
Ministry of Transport and Communications. (2018) *National Cybersecurity Strategy*. https://www.bocra.org.bw/sites/default/files/documents/approved%20botswana-national-cybersecurity-strategy.pdf
New Zealand Government. (2019) *New Zealand's Cyber Security Strategy 2019: Enabling New Zealand to Thrive Online*. https://www.dpmc.govt.nz/sites/default/files/2019-07/Cyber%20Security%20Strategy.pdf
Odebade, A.T. & Benkhelifa, E. (2023) *A comparative study of national cyber security strategies of ten nations*. *arXiv preprint* arXiv:2303.13938.
Prime Minister's Office. (2024) *Finland's Cyber Security Strategy 2024–2035*. https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/165893/VNK_2024_13.pdf?sequence=1&isAllowed=y
Republic of Argentina National Executive Branch. (2023) *Estrategia Nacional de Ciberseguridad de la República Argentina*. https://www.boletinoficial.gob.ar/detalleAviso/primera/293377/20230904
Shafqat, N., & Masood, A. (2016) *Comparative analysis of various national cyber security strategies.* International Journal of Computer Science and Information Security, 14(1), 129-136.
The Government of Japan. (2021) *Cybersecurity Strategy*. https://www.nisc.go.jp/eng/pdf/cs-senryaku2021-en.pdf
The Independent State of Papua New Guinea. (2024) *National Cyber Security Strategy*. https://www.ict.gov.pg/wp-content/uploads/2024/policies/Final%20PNG%20NCSS-%20rev%2004-04-24.pdf
The President of Uruguay. (2024) *Estrategia Nacional de Ciberseguridad del Uruguay 2024–2030*. https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/book/8671/download
President´s Office & Public Service Management and Good Governance. (2022) *Government Cybersecurity Strategy 2022.* https://www.utumishi.go.tz/uploads/documents/sw-1697717419-GOVERNMENT%20CYBER%20SECURITY%20STRATEGY%202022.pdf
The White House. (2023) *National Cybersecurity Strategy*. https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf
Weber, R. P. (1990) *Basic Content Analysis*. SAGE Publications, Inc.

Wargaming Design for Cyber Warfare

Dong Wang
New York Institute of Technology, Vancouver campus, Canada

dwang52@nyit.edu

Abstract: Wargaming design serves as a critical tool for military simulation, analysis, training, and strategic planning. Throughout history, innovations in wargaming have paralleled the introduction of new weapons and technologies. As information technology advances rapidly, cyber warfare has emerged as a fundamental component of modern conflicts. However, systematic studies of cyber warfare wargaming design remain inadequate compared to conventional kinetic warfare simulations. The cyber domain has not received appropriate attention within wargaming design frameworks, leading to conceptual confusion among terms such as "Cyber Warfare," "Cyberwar," and "Information War." This paper addresses this imbalance by synthesizing insights from traditional wargaming design and cyber warfare analysis. After examining the basic structure of traditional wargaming and assessing its adaptability to cyber warfare scenarios, this paper identified fundamental design challenges specific to cyber warfare wargaming. This research proposed an original framework based on analysis of real-world wargaming activities, particularly drawing from China's national wargaming competitions hosted by the Chinese Institute of Command and Control (CICC). The primary contributions of this research include: (1) summary and analysis of traditional wargaming structures, (2) integration of classic models with cyber warfare's unique natures, and (3) identification of key design questions requiring further studies.

Keywords: Wargaming, Cyber warfare, Wargaming design, Cybersecurity, Military simulation

1. Introduction to Wargames and Wargaming Design

Modern wargaming traces its origins to 19th century European military innovations, particularly Prussia's Kriegsspiel system. However, the fundamental philosophy underlying wargaming extends to ancient strategy games like Go and Chess, which share common principles: abstracting complex military realities into controllable simulation environments, focusing on core factors that could influence warfare outcomes, and providing insights into strategic decision-making and operational planning effectiveness, as Perla (1990) and Dunnigan (1992) stated.

Contemporary wargaming and its design methodologies rely primarily on quantitative comparisons of weapons' lethality. Carl von Clausewitz's emphasis on force concentration and massed effects in his work *On War* was reflected in modern attrition models such as Lanchester's Laws, which provide mathematical justification for firepower concentration principles through differential equations relating force size and firepower to combat outcomes. Perla (2016) discussed that in his research. These foundational concepts have guided manual and digital wargaming development for over a century.

Modern wargames are structured conflict simulations designed to train commanders, test strategies and tactics, enhance cross-service coordination, and evaluate decision-making processes within controlled environments. Their application has expanded throughout major military institutions worldwide. Several authoritative definitions of wargames exist in the literature:

In *The Complete Wargames Handbook*, Dunnigan (1992) described wargames as combining elements of gaming, history, and science, characterizing them as "glorified chess" with more complex boards and movement rules. He emphasized that wargames typically combine maps, pieces representing historical figures or military units, and rule sets governing permissible actions. Crucially, Dunnigan insisted on realism as a defining characteristic, viewing wargames as attempts to understand future possibilities through better comprehension of historical patterns.

The U.S. Department of Defense (2017) defined wargames as simulations of military operations involving opposing forces, conducted according to prescribed rules, data, and procedures for testing plans, training personnel, or developing new operational concepts in *Dictionary of Military and Associated Terms*.

In *The Art of Wargaming*, Perla (1990) defined wargames as warfare models or simulations where event flow both shapes and is shaped by human player decisions throughout the exercise. He distinguished wargames from training exercises and operations research by emphasizing humans' decision-making under uncertainty and imperfect information conditions.

By synthesizing these perspectives, this paper takes the wargame as a confrontational military simulation exercise where participants' decisions serve as the primary driving force. Well-designed wargaming possesses

predictive capabilities regarding warfare outcomes and strategic developments. Based on this understanding, wargaming design encompasses:

- Extracting and selecting key elements from real-world for scenario simulation
- Collecting and quantifying objective dynamic information, including weapons lethality, destructive capabilities, and troop movements
- Designing mechanisms that demonstrate causal relationships between decisions and outcomes

In the U.S., wargames are deeply integrated into military institutional culture and curricula. U.S. Naval War College, pioneering sophisticated manual wargames since the late 19th century, continues employing wargames as foundational tools for educating future commanders on naval strategy, operational art, and joint operations complexity, as Perla (1990) recorded. Dunnigan (1992) stated that the U.S. Army and Air Force similarly incorporate extensive wargaming, from tactical-level exercises using sand tables and computer simulations to strategy games conducted at respective war colleges.

In *Wargaming for Leaders*, Herman, Frost and Kurz (2009) viewed that regular joint wargaming activities, which involve multiple branches and international partners, are of great benefit to test joint doctrine, improve interoperability, and analyze responses to multi-domain threats. This pervasive application demonstrates U.S. military recognition of wargames as essential tools for analytical rigor, experiential learning, and adaptation to evolving operational environments.

The People's Liberation Army (PLA) of China, one of the main counterparts to the U.S. military, similarly emphasizes wargaming design as part of comprehensive military modernization. According to the U.S. Department of Defense (2023), PLA military academies and research institutions actively develop and employ advanced wargaming methodologies, recognizing demands of "intelligentized" warfare and enhanced joint operational capabilities.

The significant institutional investment in wargames by both the U.S. and China highlights their perceived critical role in preparing for modern multi-domain conflicts.

2. Cyber Warfare Conceptual Issues

2.1 Definitional Clarification

The widespread adoption of wargaming in traditional military contexts provides crucial background for understanding challenges and needs for dedicated design methodologies when applying wargaming to the characteristics of cyber warfare.

Before examining wargaming design issues, clarifying distinctions among "Cyber Warfare," "Information War," and "Cyberwar" is essential for accurate discourse and effective wargaming design in the cyber domain.

- Cyber Warfare: The U.S. Department of Defense (2017) defined this concept as the application of military force in or through cyberspace, focused on using cyber capabilities to achieve military objectives by directly affecting computer networks and systems. This represents a distinct warfare form conducted within the information domain.
- Information War (Information Operations, IO): Courterm (2024) defined IO as a broader concept than cyber warfare. In his thoughts, IO involves integrated employment of capabilities to influence, disrupt, corrupt, or usurp adversary decision-making while protecting friendly information and information systems. IO's primary objective is cognitive and behavioral influence, using various means to control or manipulate the information environment.
- Cyberwar: Hughes and Colarik (2017) found that, among 159 examined articles, only 56 provided explicit definitions of cyber war or warfare, while 103 relied on weaker implicit ones. Ashraf (2021) similarly noted the abundance of contradictory definitions—from claims of cyberwar's non-existence to warnings of its imminent threat—and proposed instead a flexible framework based on three core themes and five variables drawn from an interdisciplinary literature review. Dyer-Witheford and Matviyenko (2019) stated that cyberwar is often used imprecisely in public and media contexts to describe significant hostile cyber activity between states. This encompasses various actions, including espionage, sabotage, and disruption.

2.2 The Gap in Cyber Warfare Wargaming Design

After clarifying cyber warfare scope, a significant gap emerges: few wargames focus primarily on cyber warfare. Most existing military exercises treat cyber as service or support elements rather than primary focus areas. Several factors contribute to this deficiency:

- Abstract and Intangible: Unlike physical assets such as vehicles or weapons, cyber "assets" and "casualties" lack physical form. Traditional wargames employ quantitative models to measure force exchange; however, no analogous, agreed-upon metrics exist for cyber effects. The absence of clear attrition ratios for cyber forces complicates deterministic outcome assignment to cyber engagements.
- Organizational Integration: Cyber units are often organized as support elements for conventional forces, with roles that may not be independently modelled.
- Lack of Assessment Consensus: Cyber warfare commentators range from viewing cyber weapons as strategic nuclear threat equivalents requiring extreme caution to dismissing them as minor nuisances if physical infrastructure can be isolated, as reported in *War in the Fifth Domain* (2010). This divergence between "cyber hype" and "cyber scepticism" complicates scenario design.

The shortage of dedicated cyber wargaming design matters because modern conflicts already incorporate cyber dimensions. For non-state actors or rogue state, cyber warfare offers excellent cost-effectiveness ratios. These actors can conduct cyberattacks without physical exposure, explaining why certain authoritarian states invest heavily in cyber warfare technologies. The complexity, persistence, and concealment characteristics of cyber warfare mean that discovery of cyberattacks often occurs too late for effective response, necessitating advance consideration and response planning.

3. Integrating Cyber Warfare into Wargaming Design Paradigms

Traditional wargaming frameworks provide valuable structural foundations, but their components require thoughtful adaptation to effectively represent cyber warfare's characteristics.

3.1 Structural Adaptations

3.1.1 Scenario

Scenarios should represent realistic cyber warfare situations, drawing inspiration from historical cyber conflicts, anticipated future threats, or specific operational plans. Traditional wargaming scenarios require modification to represent multi-faceted cyber warfare, which may involve system degradation, data compromise, information manipulation, or service denial rather than purely physical destruction.

3.1.2 Map/board representation

Traditional wargaming philosophy stimulates real battle situations in controlled environments, typically represented as "maps" or "boards." In contrast, cyber warfare wargaming "maps" represent complex, interconnected network landscapes including:

- End users (laptops, servers, drones, mobile devices)
- Protection measures (firewalls, intrusion detection systems, security policies)
- Critical infrastructures (channels, cables, power grids, financial networks, communication systems)
- Service providers (data centres, cloud computing platforms)
- Emerging AI infrastructures

3.1.3 Units/pieces

Traditional wargaming "pieces" represent weapon units or battlefield events. Quantified points assigned to these pieces provide fundamental elements for evaluating battle outcomes.

In cyber warfare wargaming, each “unit” should be defined by specific capabilities, limitations, and potential rule-based interactions, representing various entities impacting cyber environments:

- Offensive cyber units (network attack technologies and exploitation capabilities)
- Defensive cyber units (network defense systems, incident responses, and resilience capabilities)
- Critical infrastructures (modelled with specific vulnerabilities and dependencies)
- Data assets (with varying sensitivity and value levels)
- Representations of non-state actors or individual users serving as attack vectors or influence targets

3.1.4 Turn sequence/rounds

Sequential turn structure in traditional wargaming is a necessary trade-off between reality and time management practicality. However, it is inadequate in the context of rapid and simultaneous cyber warfare scenarios. The design of turn sequences of cyber warfare wargaming must account for speed and concurrency of events across distributed networks. Options include highly abstracted simultaneous turns, discrete phases allowing specific action types, or real-time/near-real-time adjudication facilitated by control teams or automated mechanisms.

3.1.5 Force status

In cyber warfare scenarios, cyberattack weapons achieve maximum effectiveness when unveiled to opponents (e.g., Zero-day exploits, Advanced Persistent Threats, System backdoors). Once deployed, the openness and versatility of cyberattacks enable opponents to quickly patch or address vulnerabilities. This characteristic necessitates rapidly changing parameter settings in the design of force status.

Cyber warfare wargaming's force tracking must watch not just physical loss, but whether critical nodes are exploited, controlled or denied, and how C.I.A. triad (Confidentiality, Integrity and Availability) degrades. It also has to keep an eye on network performance, defensive-tool readiness and the cascading side-effects that flash through linked systems in seconds.

3.2 Key Questions for Designing

3.2.1 What are the cyber warfare wargaming's objectives?

The foundational question for any wargaming design concerns its purpose. In cyber warfare, objectives are particularly diverse, ranging from technical training to strategic analysis. Chosen objectives dictate required technical fidelity levels, simulated environment scale, game duration, and participant expertise requirements. Clearly defining analytical or training objectives is paramount before other design decisions.

3.2.2 How is victory determined?

Defining victory in cyber warfare poses significant challenges compared to traditional kinetic conflicts. Whyte and Mazanec (2019) highlighted that cyber attacks' outcomes are often intangible, delayed, indirect, and difficult to attribute definitively. Successful attacks typically yield data exfiltration, system degradation, or information manipulation rather than physical destruction. This renders conventional metrics like attrition or territorial gains largely inapplicable. The paper discusses combined victory criteria tailored to cyber warfare.

3.2.3 How to model cyber domains?

Unlike relatively stable physical wargame terrain, cyber domains are constantly in flux. New vulnerabilities are discovered, exploits are developed, patches are released, and technologies rapidly evolve. Cyber warfare wargaming design should realistically represent this dynamism through the control team's intervention or AI-powered automated mechanisms.

3.2.4 How to represent third-party participants?

Cyber warfare involves diverse actors beyond traditional state militaries, including intelligence agencies, non-state proxies, criminal organizations, and ideologically motivated groups. These actors possess varying capabilities, motivations. Effective wargaming design should represent this complex ecosystem accurately.

3.2.5 How to incorporate legal and policy constraints?

Cyber warfare occurs within complex and ambiguous legal and policy frameworks, including international laws, national rules of engagement, and evolving international cyberspace behavioral norms. Modelling and designing these non-technical factors are essential for strategy wargaming because they can authentically reflect the complexities of real-world decision-making procedures.

4. Case Study: Chinese Modern Wargaming Design

4.1 CICC 2025 Wargaming Competition

Since 2016, the Chinese Institute of Command and Control (CICC) has hosted annual national wargaming competitions from May to December, with participants primarily from military academies and research institutions. The ongoing 2025 competition comprises five events:

Table 1: Basic Information about 2025 CICC Wargaming Competitions

Competition Category	Wargaming Level	Platform Used	Branches Involved	Wargaming Design Focuses On
People vs. People Wargaming Match	Strategy joint operation	"Mozi Future Commander" wargaming system	All military branches	Joint confrontation across land, sea, air, and cyber-electronic warfare domains based on geopolitical conflicts
Joint Operations People-System Collaborative Challenge	Campaign-level joint operations	"Lingyi (Smart Gaming)" intelligent wargaming platform	All military branches	People-System collaboration models and decision-making technologies in large-scale, complex confrontations
Intelligent Air Combat Algorithm Challenge	Action-level	"Zhikong (Seize Sky)" air combat wargaming platform	Air force	Intelligent gaming algorithms for unmanned systems
Urban Operations Algorithm Challenge	Action-level	"Chengchi Gonglue (City Capture)" wargaming platform	Army	Decision-making assistance and gaming confrontation technologies
Long-Range Collaborative Strike Offensive and Defensive Challenge	Campaign-level joint operations	"Juesheng Qianli" (Winning From a Thousand Miles Away) intelligent wargaming platform	Navy	Large model intelligent cognition and dynamic gaming in full-process, long-range collaborative strike scenarios
Traditional Board Wargaming Contest	Educational event	"Red Classics" board-and-turn-based wargames	Army	Historical battles on Chinese front during World War II

The nine-year-old competitions have significantly supported Chinese military development, including the formations of Chinese Cyberspace Force and Information Support Force, which have made their debuts in September's V-Day military parade through Tian'anmen Square.

Except for the educational "Traditional Board Wargaming Contest," all competitions utilize specially designed network platforms with client-server architecture, on which software platforms are deployed and operated. These platforms support daily training and exercises.

4.2 CICC 2024 Competition Example

Using CICC 2024 final competition as an example, typical turn-based wargaming processes include the following components.

Table 2: The Final Competition Scenario of CICC 2024 is Named "Pincer Movements at Kalinin"

Components	Main Contents/Purposes	Scenario Description
Wargaming Purpose	Outlines principal assessment criteria, design features, core contested issues, and weapons peculiarities, enabling the control team to evaluate participants' strategic planning performance	Exercise enabling commanders to conduct strategic planning within multi-domain joint operations, integrating land, maritime, air, and cyber-electronic warfare assets
Participants	In competitive wargaming, Red and Blue cells form small teams; each designates one overall commander while others serve as operators executing orders and managing force allocation	Red cell and Blue cell (three personnel each); A control team serves as coordinator and adjudicator
Wargaming Background	To maximize real-world utility, wargaming focuses on current hot-spot regions and ongoing crises, testing participants' strategic decision-making in authentic conflicts	Y2028: Buffer-state B tilts to R-State; Atlantic-led Alliance strikes first to keep the east out of Europe's doorstep, yet its own northern door is ajar
Wargaming Period	Select plausible conflict scenarios emerging in near-future timeframes	04 OCT 2028, 10:00:00–12:30:00 Zulu
Win/Loss Conditions	Victory assessed by quantified points only—destroy opponent's weapon power and minimize own losses; personnel casualties not factored	Destroy key target: attacker gains points, defender loses; highest point at end wins
Map	Interactive digital maps of physical battlefields	European-Russian border area, annotated per scripted force lay-down and territorial ownership

Components	Main Contents/Purposes	Scenario Description
Force Status	Provide separate Red and Blue weapon equipment lists with assigned power points	Total points for each side balanced within ±1.5%
Mission Briefings	Primary directive for joint-force exercise; each participant's deployments and strategic planning are solely under its authority	By order of Red/Blue Combined Joint Force HQ, the operational directives outlined the mission background, opponent's situation, assigned tasks, and timing requirements

5. Wargaming Design Framework for Cyber Warfare

Based on CICC past year's wargaming analysis, a framework for cyber warfare wargaming design can be proposed. The framework addresses the unique characteristics of cyber warfare while maintaining the proven structure of traditional classic wargaming methodology.

5.1 Scenario Design

Given the rapid iteration of cyber technologies, the time span of a wargaming should not be too long, preferably measured in monthly units, with the scenario set for a few months into the future. This timeframe allows for realistic technological evolution while avoiding speculation about distant future capabilities. The introduction of third-party variables can reflect the complexity of cyber warfare, including the actions of non-state actors, criminal organizations, and proxy groups that often complicate real-world cyber conflicts.

Scenarios should be grounded in current threat landscapes and emerging technological trends. They should incorporate realistic geopolitical tensions and reflect the interconnected nature of global information infrastructure. The scenario should also account for the civilian nature of much cyber infrastructure, including the role of private companies as both targets and defenders.

5.2 Map Design

Instead of a physical battlefield, a network topology map should be the primary medium to reflect the unique characteristics of cyber warfare. This map should represent:

- Critical infrastructure nodes (power grids, financial systems, communications networks)
- Government and military networks with varying classification levels
- Commercial networks and cloud service providers
- Internet service providers and routing infrastructure
- Mobile and IoT device networks
- International connections and undersea cables

The map should be dynamic, allowing for changes in network topology as systems are compromised, isolated, or restored. It should also represent logical connections rather than just physical geography, showing data flows and trust relationships between different network segments.

5.3 Parallel Operations Design

To address the problem of synchronized realism, many cyber exercises shorten the time slices of each turn to simulate real-time synchronization, but they are fundamentally still asynchronous and turn-based. However, cyber warfare wargaming design should be strategically and tactically oriented rather than technically driven. It's recommended to introduce a comprehensive adjudication mechanism in the figure below.

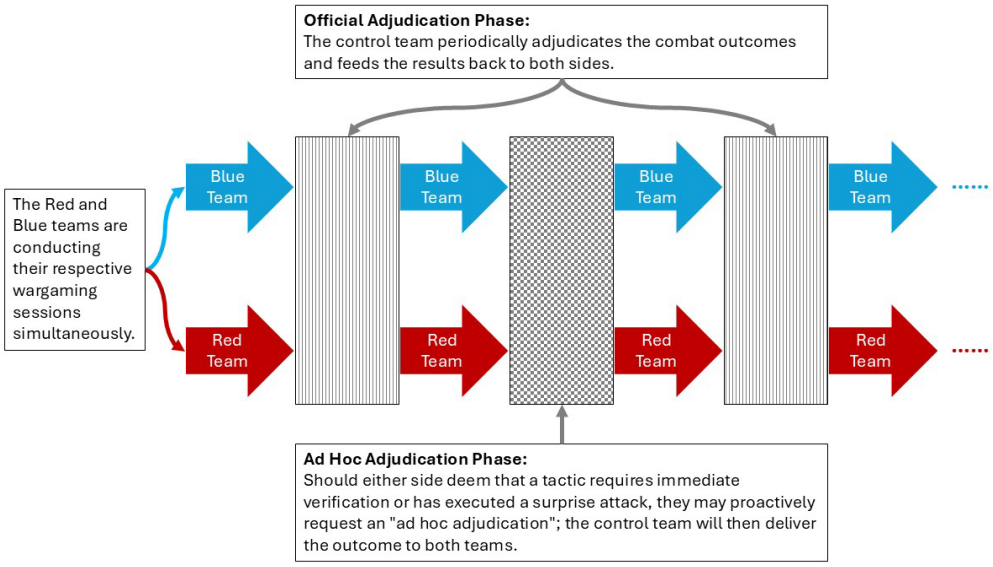


Figure 1: A New Design Adjudication Mechanism for Cyber Warfare Wargaming

This mechanism foregrounds the function of the control team, integrating it into the wargaming process. It safeguards the uninterrupted continuity of Red and Blue cells' strategic planning while simultaneously accommodating flexibility and complexity. The mechanism includes the following main phases:

- Planning and Execution Phase: Red and Blue teams carry out their moves simultaneously, guided by their mission orders and their interpretation of the wargame settings.
- Official Adjudication Phase: At intervals defined by the scenario, the control team assesses the battlefield situation and feeds the results back to all parties. Note that the time between two adjudications is not fixed—the unit of time is variable and may be denominated in hours, days, or even in seconds and milliseconds.
- Ad-Hoc Adjudication Phase: At any moment, either side may request the control team to intervene based on its strategy or operational developments; the ruling is then relayed to all parties.

5.4 Win/Loss Conditions Design

Cyber warfare wargames differ significantly due to the existence of sudden-death nodes. The destruction of certain nodes, such as undersea cables or major data centres, can cause entire networks to collapse. Therefore, the win/loss conditions should include both the cumulative calculation of destroyed enemy cyber forces and the potential for a sudden-death attack.

This design forces decision-makers to train under the pressure of a sudden-death threat while balancing multiple competing objectives, further highlighting the wargaming training value.

5.5 Force Status Design

In traditional wargaming, the quantitative points of weapons are often fixed. However, in cyber warfare wargaming, the power of a cyberattack weapon and the resilience value of a cyber-defense weapon are variable. Since a change at one node can alter the entire network topology, the force status should change dynamically under different offensive and defensive conditions. The force status system should track:

- Offensive powers' changing status (e.g., Zero-day exploits and their degradation after disclosure)
- Defensive capabilities and their effectiveness against known attack vectors
- System compromise levels (ranging from reconnaissance to full control)
- Network robustness
- Attribution confidence levels for observed activities

5.6 Third-Party Participants Design

In addition to the standard Red vs. Blue cells confrontation, the introduction of third-party participants increases the diversity and randomness of the wargame. These third parties might include:

- Criminal organizations conducting attacks for financial/political rewards
- Hacktivist with ideological motivations
- Private companies providing cyber defense services
- Allied nations with their own cyber capabilities and objectives
- International organizations attempting to mediate or investigate

5.7 AI Integration Design

AI itself is part of the substance of cyber warfare wargaming. Whether the models are secure, reliable and controllable constitutes an unknown risk for every participant. Learning to strike the right balance in employing AI tools is a training objective in the wargaming context.

On the other hand, the integration of AI into cyber warfare wargaming could reduce the calculation burden on the control team. AI can perform as third-party participants with realistic but unpredictable behavior patterns. Furthermore, AI can dynamically adjust force status points according to the always-changing network topology, which can enrich the scenario and bring more flexibility.

AI-integration is the issue that needs careful reflection and cautious handling. The author is also investigating the root causes of so-called AI "ghost answers" or fabricated responses. One potential factor is contamination in the datasets used to train these models. For example, a study led by Alexandra Souly (2025) demonstrates that Poisoning Attacks, in which adversaries inject malicious documents into training datasets, can undermine the safety and reliability of large language models.

6. In Conclusion

This paper sets out a cyber warfare wargaming design framework that mirrors cyber domain's special traits. It tries to make breakthroughs and innovations in four aspects. First, introducing third-party participants multiplies the complexity of strategic planning. Second, a parallel operation mechanism allows any participant to call for ad-hoc adjudication at certain moments, and this design can simulate real situations as much as possible. Third, a sudden-death mechanism widens victory beyond simple point calculation to a triad: destroy enemy capabilities, protect critical assets, and create windows for surprise cyber strikes. Fourth, network-topology maps replace physical battlefield charts, presenting cyber warfare features more scientifically.

Future work should proceed along three main directions: (1) Empirical validation involves systematically collecting and analyzing winning CICC wargaming after-action reports. (2) Comparable archives from other wargaming events, to assess the effectiveness of the proposed design. (3) Building on the ideas presented in this paper, the author is currently developing a cyber warfare wargaming platform that applies these design principles in practice, aiming to enhance experiential learning and operational understanding in cyber warfare, cybersecurity, and broader network-based operations.

In the context of cyber warfare wargaming, humans' decision-making remains irreplaceable; AI should only be assigned narrow, tightly controlled tasks—such as acting as third-party participants, supporting the control team's adjudications or enriching wargaming scenarios—and must never be allowed to encroach on strategic planning.

AI Declaration: No AI tools were used in the development of this paper.

Ethics Declaration: This research did not require ethical approval.

References

Ashraf, C. (2021) 'Defining cyberwar: towards a definitional framework', Defense & Security Analysis, 37(3), pp. 274–294. doi:10.1080/14751798.2021.1959141.

Boer, L.J.M. (2021) International law as we know it: cyberwar discourse and the construction of knowledge in international legal scholarship, Cambridge: Cambridge University Press.

Carr, J. and Shepherd, L. (2010) Inside cyber warfare, Sebastopol, CA: O'Reilly.

CICC Organizing Committee (2024) "**关于**组织全国兵棋推演大赛人人对抗主体赛比赛平台公测的通知 [Notice on the Public Testing of the Wargaming Platform for the People vs. People Wargaming Match of the National Wargaming Competitions]" (in Chinese), [online], Beijing: CICC, https://www.ciccwargame.com/h-nd-135.html (Accessed: 18 September 2025).

CICC Organizing Committee (2024) "**关于**举办 2024 **第八届全国兵棋推演大**赛 **人机混合决策**专项赛的通知 [Notice on the 8th National Wargaming Competitions 2024 – People-system Hybrid Decision-Making Special Competitions]" (in Chinese), [online], Beijing: CICC, https://www.ciccwargame.com/h-nd-141.html (Accessed: 18 September 2025).

CICC Organizing Committee (2025) "**关于**举办 2025 **第九届全国兵棋推演大**赛的通知 [Notice on the 9th CICC National Wargaming Competitions 2025]" (in Chinese), [online], Beijing: CICC, https://www.ciccwargame.com/h-nd-146.html (Accessed: 18 September 2025).

Cunningham, C. and Touhill, G. (2020) Cyber warfare: truth, tactics, and strategies – strategic concepts and truths to help you and your organization survive on the battleground of cyber warfare, 1st edn, Birmingham, UK: Packt Publishing.

Death, D. (2017) Information Security Handbook: Develop a Threat Model and Incident Response Strategy to Build a Strong Information Security Framework, Birmingham, UK: Packt Publishing.

Dunnigan, J.F. (1994) "Complete wargames handbook: how to play, design, & find them (rev ed)", Airpower Journal, 8(2), pp. 85–87.

Dyer-Witheford, N. and Matviyenko, S. (2019) Cyberwar and revolution: digital subterfuge in global capitalism, Minneapolis, MN: University of Minnesota Press.

Hughes, D. and Colarik, A. (2017) 'The Hierarchy of Cyber War Definitions', in G. Wang, M. Chau and H. Chen (eds) Intelligence and Security Informatics: PAISI 2017. Lecture Notes in Computer Science, vol. 10241. Cham: Springer, pp. 19–29. doi:10.1007/978-3-319-57463-9_2.

Libicki, M.C. and Project Air Force (2009) Cyberdeterrence and cyberwar, Santa Monica, CA: RAND Corporation.

Lonergan, E.D. and Lonergan, S.W. (2023) Escalation dynamics in cyberspace, New York: Oxford University Press.

Marshall, J. (2009) "Wargaming for Leaders: strategic decision-making from the battlefield to the boardroom", Financial Executive, 1 January.

Office of the Chairman of the Joint Chiefs of Staff (2017) "DOD dictionary of military and associated terms", [online], Washington, DC: The Joint Staff, https://apps.dtic.mil/sti/pdfs/AD1029823.pdf (Accessed: 18 September 2025).

Perla, P.P. (1990) The art of wargaming: a guide for professionals and hobbyists, Annapolis, MD: Naval Institute Press.

Perla, P.P. (2016) "Operations Research, Systems Analysis, and Wargaming: Riding the Cycle of Research", in Harrigan, P. and Kirschenbaum, M.G. (eds.) Zones of control: perspectives on wargaming, Cambridge, MA: MIT Press.

Poindexter, D.F. (2015) The new cyberwar: technology and the redefinition of warfare, Jefferson, NC: McFarland & Company, Inc., Publishers.

Shakarian, P., Shakarian, J. and Ruef, A. (2013) Introduction to cyber-warfare: a multidisciplinary approach, 1st edn, Amsterdam: Morgan Kaufmann Publishers, an imprint of Elsevier.

Souly, A. et al. (2025) 'Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples'. Available at: https://research.ebsco.com/linkprocessor/plink?id=c10adda0-caaa-3e6a-b9da-1b014b6aadf4 (Accessed: 18 December 2025).

The Economist (2010) "War in the fifth domain: Cyberwar", The Economist, 3 July.

U.S. Department of Defense (2023) "Military and security developments involving the People's Republic of China", [online], Washington, DC: Department of Defense. https://media.defense.gov/2023/Oct/19/2003323409/-1/-1/1/2023-MILITARY-AND-SECURITY-DEVELOPMENTS-INVOLVING-THE-PEOPLES-REPUBLIC-OF-CHINA.PDF (Accessed: 18 September 2025).

Weng, M.-H. and Chang, H.-C. (2019) **兵棋推演**：**意涵**・**模式與操作** [Wargaming: meaning, models, and practice], 1st edn, Taipei: Wu-Nan Book Inc. (in Chinese).

Whyte, C. and Mazanec, B.M. (2023) Understanding cyber warfare: politics, policy and strategy, 2nd edn, London: Routledge, Taylor & Francis Group.

Winterfeld, S. (2013) The basics of cyber warfare: understanding the fundamentals of cyber warfare in theory and practice, Boston: Syngress.

Yang, N. (ed.) (2007) **虚**拟演兵：兵棋、作战模拟与仿真 [Virtual wargaming: wargame, combat simulation and emulation], Beijing: **解放**军出版社 [PLA Publishing House]. (in Chinese).

Non Academic Papers

A Review of the Maritime Cybersecurity Regime Over the Last Decade

Leigh Armistead
Peregrine Technical Solutions, Juneau, Alaska, USA

leigh.armistead@goldbeltfed.com

Abstract: This paper discusses the development of how the Maritime Cybersecurity regime has evolved over the last decade due to new instructions, mandates, doctrine and criteria that has been established to better protect both the traditional Internet Technology (IT) assets as well as the Industrial Control Systems (ICS). The methodology of this paper is to layout the new regulations over the last 10 years, and to showcase how these mandates affect operational issues at sea. The key factor, that will pervade all of the discussions of cybersecurity in the maritime domain, is the intense focus on safety, only to be followed by availability. This is unique, as the IT sector traditionally is focused on confidentiality, but Operational Technology (OT) or Facility Related Control Systems (FRCS) are a different set of systems, which require continuous operation, hence the need to always be on. In addition, the author will lay out in a systematic function, all of the new regulations in the last 10 years, why they were needed, and how did they change the training, skillsets, as well as processes that have been implemented to ensure compliance moving forward. Peregrine Technical Solutions, LLC. (Peregrine) is led by Dr Leigh Armistead, CISSP, who has conducted a number of maritime cybersecurity same tasks since 2014. It is a small business, based in Juneau, Alaska (US), specializing in ICS/OT/FRCS cybersecurity, where we have successfully conducted assessments over the last decade for the US Department of Defense (DoD) as well as academia. We are currently the prime contractor for the US Coast Guard Advanced Metering Systems, and we won an Army contract in 2023 for Facility Related Control Systems (FRCS) cybersecurity. From 2017-2021, we were the lead for the Platform Resilience Mission Assurance (PRMA) for the DoD, plus we have the first Cybersecurity Department of Labor Registered Apprenticeship Program (RAP) in the nation for adults (2016) /Youth Registered Apprenticeship (YRA) (2019). Dr Leigh Armistead was a member of the NATO project, SAS-163, Energy Security in the Era of Hybrid Warfare. This paper is a series of case studies of contracts and research efforts that Peregrine staff conducted over the last decade. It lays out a series of scenarios, using action research, as part of a qualitative methodology, to demonstrate the changes for the maritime community from a cybersecurity aspect. As this is a series of operational actions, there is not a literature review as such but instead, we abide by regulations, mandates, instructions and notices that are issued by a variety of regulatory bodies and organizations. The key participants are as follows: International Maritime Organization (IMO) – Regulatory Body. Peregrine Technical Solutions – FRCS Cybersecurity Provider. Academic Research Fleet (ARF) – 18 Ships. National Science Foundation (NSF) – Contract Sponsor. University of California San Diego (UCSD) – Original Organization requesting Support. Woods Hole Observatory Institute (WHOI) – Follow-on Contract Sponsor

Keywords: Cybersecurity, Maritime, ICS, FRCS, OT

1. Initial Cyber Security Efforts

In June 2017, the International Maritime Organization (IMO) released a Maritime Safety Committee (MSC) resolution that addresses Maritime Cyber Risk Management in Safety Management Systems (MSC-FAL.1/Circ. 3). This resolution "affirms that an approved safety management system should take into account cyber risk management in accordance with the objectives and functional requirements of the ISM Code". The resolution also encourages administrations to "ensure that cyber risks are appropriately addressed in safety management systems no later than the first annual verification of the company's Document of Compliance after 1 January 2021." This was the first time that ships and their crew were directed to abide by a cybersecurity requirement, and it gave the community over three years to be ready to meet this deadline. Of note, these new requirements were embedded as a safety issue, which meant they were not ignored (as much), and needed to be met to pass the inspection.

Peregrine was requested, starting in 2018 to help the University of California of San Diego (UCSD) to meet this requirement. Over a two-year period, these subject matter experts (SMEs) developed a series of recommendations and in October 2020, a contract was let from the National Science Foundation (NSF) to support all 14 Universities of the Academic Research Fleet (ARF) which is comprised of 18 ships across the United States. We started a series of cybersecurity efforts on 20 November 2020 and conducted over a year long period to satisfy the MSC-FAL.1/Circ. 3 mandate, based on these key elements below:

- Identify: Define personnel roles and responsibilities for Cyber Risk Management (CRM) and identify the systems, assets, data, and capabilities that, when disrupted, pose risks to ship operations.
- Protect: Implement risk control processes and measures, and contingency planning to protect against a cyber-event and ensure continuity of shipping operations.

- Detect: Develop and implement activities necessary to detect a cyber-event in a timely manner. Update procedures for reporting non-conformities, accidents, and hazardous situations to include reports relating to cyber incidents.
- Respond: Develop and implement activities and plans to provide resilience and to restore systems necessary for shipping operations or services impaired due to a cyber-event. Ensure that adequate resources and shore-based support are available to support the Data Protection Act in responding to the loss of critical systems.
- Recover: Identify measures to back-up and restore cyber systems necessary for shipping operations impacted by a cyber-event. Include creation and maintenance of back-ups into the ship's operational maintenance routine.

Our team conducted at the height of COVID, a robust and comprehensive cybersecurity effort where we gave bonus support not requested in the original RFP, specifically for the six Navy-owned ships, that had immediate Defense Federal Acquisition Regulation (DFARS) 252.204-7012 requirements, where we supplied the 10 required files, plus conducted table-top exercises with the staffs at Woods Hole Observatory Institute (WHOI), University of Hawaii (UH), University of California San Diego (UCSD) and University of Washington (UW). Our customer was extremely happy that we went above and beyond the contracted deliverables by doing things such as delivering detailed HW/SW inventories for each ship and attending American Bureau of Shipping (ABS) inspections virtually, so that we could answer all questions regarding cyber security for each ship. On that original 2020-2021 effort, from our interviews with the staff of the 14 Universities and there was a wide variety of knowledge and expertise. Some of the larger Universities, had a relatively strong cyber security posture, with no Information Technology (IT) / Operational Technology (OT) overlap meaning that it is unlikely that a threat actor could damage ship operations remotely. On others, not so much, as we found that there is a large amount of network segmentation on the ships from the larger Universities, where science equipment is separated from other devices. And in virtually instances, the ships crew did not have a current list of all of their equipment, which we provided as part of this contract.

The goal of these assessments was to analyse and assess the ship's network, its systems/devices to identify any vulnerabilities that could compromise or result in either loss of confidentiality, loss of integrity, or result in a loss of operation of the equipment, system, network, or even the ship. These vulnerabilities and weaknesses could fall into one of the following categories:

- Technical – software defects, End of Life (EOL) or End of Support (EOS) systems
- Design – access management, unmanaged network interconnections
- Implementation – errors such as misconfigured firewall rules or access control lists
- Procedural – system and physical access or other user errors

In addition, for this effort, Peregrine cybersecurity subject matter experts (SMEs) conducted these cyber ship assessments, in which the following items were completed:

- Map the ship's key functions and systems and assign potential impact levels using the Confidentiality, Integrity, and Availability (CIA) model, taking into consideration the operation of OT systems
- Identify main producers of critical shipboard IT and OT equipment
- Review detailed documentation of critical OT and IT systems including their network architecture, interfaces, and interconnections
- Identify cyber security points-of-contact with each of the producers and establish a working relationship with them
- Review detailed documentation on the ship's maintenance and support of the IT and OT systems
- Establish contractual requirements and obligations that the shipowner/ship operator may have for maintenance and support of shipboard networks and equipment
- Support the risk assessment with an external expert to develop detailed plans, including producers and service providers as necessary

2. Different Focus from Traditional Ashore IT Cyber Assessments

From a Maritime Cyber Security Implementation aspect, historically the ongoing, day-to-day operational efforts focused on availability, and to identify cybersecurity vulnerabilities for its potential impact and the probability of its exploitation immediately. Shown below are the recommended technical and/or procedural

corrective actions at a high-level that should be used, and ideally, there should be a cyber risk assessment for each identified vulnerability:

- Executive summary – a high-level summary of results, recommendations, and the overall security profile of the assessed ship
- Technical findings – breakdown of discovered vulnerabilities, their probability of exploitation, the resulting impact, and appropriate technical fix and mitigation advice
- Prioritized list of actions – the priorities allocated should reflect the effectiveness of the measure, the cost, the applicability, etc. It is important that this list should be a complete list of options available and not represent a list of services and products the third-party risk assessor, if applicable, would like to sell.
- Supplementary data – a supplement containing the technical details of all key findings and comprehensive analysis of critical flaws. This section should also include sample data recovered during the penetration testing, if any, and all critical or high-risk vulnerabilities.
- Appendices – records of activities conducted by the cyber risk assessment team and the tools used during the engagement.

As alluded earlier, the ARF engagement allowed the entire fleet to pass its yearly ABS inspection for cybersecurity over a four-year period starting in 2021. The activities performed during this previous onsite ship assessment included reviewing the configuration of all computers, servers, switches, routers, and other network enabled technologies, plus cyber security procedures and documentation for connecting and managing all available IT and OT systems and devices were also reviewed. In addition, the Peregrine SMEs also conducted a virtual assessment of the involvement of the crew of all levels; particularly the master, chief engineer and first mate. By doing this, the Peregrine team could better understand the implementation of the IT/OT systems onboard and how they may vary from stated or recommended design documentation. This assessment methodology also helped in understanding the level of cyber training delivered to the ship's crew.

Specifically, the Peregrine staff conducted passive IT/OT network scans, storing all data in a secure portal, specific to WHOI. Data collection devices were scrubbed after the data was transferred securely to the backup device so as to avoid any leakage of confidential data. Information and findings were transferred from the IMO Test Plan Template to the implementation working documents:

- Technical Findings Worksheet Template
- Supplementary Data Worksheet Template - Supports technical findings
- Appendices - Authoritative documents, requirements cited
- Findings were analyzed, and risks identified for mitigation based on criticality to personnel, operations, Confidentiality/Integrity/Availability (CIA) values, and mission assurance
- Customer Relationship Management (CRM) analysis performed; triaged risks identified in Risk Sheet Template
- Actions Worksheet Template sets out the outbound actions required to mitigate risk
- Management Implementation Begins

A key element of all sites was to understand the H/W and S/W of the various ships, and the Peregrine cybersecurity SMEs, in which the following items were completed:

- Map the ship's key functions and systems and assign potential impact levels using the Confidentiality, Integrity, and Availability (CIA) model, taking into consideration the operation of OT systems
- Identify main producers of critical shipboard IT and OT equipment
- Review detailed documentation of critical OT and IT systems including their network architecture, interfaces, and interconnections
- Identify cyber security points-of-contact with each of the producers and establish a working relationship with them
- Review detailed documentation on the ship's maintenance and support of the IT and OT systems
- Establish contractual requirements and obligations that the shipowner/ship operator may have for maintenance and support of shipboard networks and equipment
- Support the risk assessment with an external expert to develop detailed plans, including producers and service providers as necessary

The goal of these assessments was to analyse and assess the ship's network, its systems/devices to identify any vulnerabilities that could compromise or result in either loss of confidentiality, loss of integrity, or result in a loss of operation of the equipment, system, network, or even the ship. These vulnerabilities and weaknesses could fall into one of the following categories:

- Technical – software defects, End of Life (EoL) or End of Support (EoS) systems
- Design – access management, unmanaged network interconnections
- Implementation – errors such as misconfigured firewall rules or access control lists
- Procedural – system and physical access or other user errors

3. Follow-on Maritime Cyber Security Implementation

After this original one-year contract, we were engaged again by one of the larger Universities in the ARF in 2024 with a Statement of Work (SOW) for Cyber Security Assessment Support to meet the new American Bureau of Shipping (ABS) Guide for Cybersecurity Implementation for the Marine and Offshore Industries (ABS CyberSafety Volume 2) dated August 2023. Peregrine conducted two Site visits to vessels in both the Pacific and Atlantic Oceans that were conducted in October – November 2024 time frame, and based upon those surveys, the Peregrine staff evaluated both vessels with CS-1 status, based on the 40 requirements per sections 1.3 of the American Bureau of Shipping (ABS) CyberSafety Vol 2, pages 26-31, as well as section 2.3 (page 34). In addition, Peregrine SMEs also reviewed with our findings with both ship's crew, to confirm that they understood and were abiding by the ABS Appendices shown below:

- Appendix 1 Maritime Cybersecurity Risk Assessment
- Appendix 2 Functional Description Document (FDD)
- Appendix 3 Cybersecurity Risk Management System (CRMS)

In addition, since Peregrine is a sponsor of the Maritime Transportation System (MTS) Security Operations Center (SOC) since 2019, https://www.mtsisac.org/, we gave for this effort the University access to all Maritime cybersecurity threats from this Department of Homeland Security (DHS) sponsored Information Sharing and Analysis Center (ISAC).

For Peregrine's most recent cyber assessments, we based it on a number of references to include:

- The Guidelines on Cyber Security Onboard Ships, version 4 published 23 December 2020 (hereafter referred to as the Baltic and International Maritime Council (BIMCO) document)
- MSC-FAL.1/Circ. 3 Guidelines
- The ABS CyberSafety Volume 2) dated August 2023

Our SMEs worked with the respective ship's crew as well the Universities IT staff to finalize the files for submission to ABS. To do this, our team followed the specific guidance from the three aforementioned references, which makes clear that an approved SMS should consider cyber risk management when meeting the objectives and functional requirements. These three documents also provided guidance on maritime cyber risk management as well as high-level recommendations regarding the elements of an appropriate approach to implementing cyber risk management. This document is designed to provide the minimum measures that all companies should consider implementing to address cyber risk management in an approved SMS.

Based on The Guidelines on Cyber Security Onboard Ships, version 4 published 23 December 2020 (hereafter referred to as the BIMCO document), along with the MSC-FAL.1/Circ. 3 guidelines, and the ABS CyberSafety Volume 2) dated August 2023, these three documents would serve as the primary reference for the Peregrine SMEs to review for their latest 2024-2025 cybersecurity accreditation effort.

- Company and Vessel Cybersecurity Representative or Organization (1 August 2023) - Company to provide documentation identifying person(s) responsible for cybersecurity, the organizational location within the Company, and quality or cybersecurity certificates (e.g., International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) 27001, International Society of Automation (ISA)/IEC 62443, or ISASecure® certifications).
- Company and Vessel Cybersecurity Policies and Procedures (1 August 2023) - Company to provide documentation detailing cybersecurity policies and implementation procedures applied to employees, suppliers, and contractors.

- Company and Vessel Cybersecurity Incident Response (IR) and Recovery Documentation describing Company and vessel cybersecurity incident response team(s), including response and recovery responsibilities, capabilities, and staffing.
- Company and Vessel OT and Related IT System Architecture - Descriptions and diagrams detailing the OT and connected IT systems architecture that are suitable for performing a physical and cybersecurity risk assessment.
- Company and Vessel OT and Related IT Risk Assessment and Management Plan - A cybersecurity risk assessment of the OT and connected IT systems architecture, and a risk management plan derived from that assessment.
- Company and Vessel Cybersecurity Risk Management System (CRMS) Design, Operating and Maintenance Procedures - CRMS functional description document (FDD) that includes a functional architecture with an inventory and descriptions of cybersecurity protective equipment; logical and procedural protections (see Appendix 3); operating and maintenance procedures; cybersecurity event/incident corrective/ preventive action procedures; auditing and test procedures; and, change management procedures.
- Company and Vessel Cybersecurity Training Program - Documentation detailing cybersecurity training records and training content concerning cyber hygiene and support of specialized cybersecurity functions.
- Company and Vessel OT/IT Management of Change (MOC) Procedures - Documentation detailing change management and configuration control policies and procedures applied during vessel operation, including OT, OT-connected IT, and CRMS software and computer hardware registries

4. Recommendations Moving Forward

Here are some specific suggestions from the Peregrine staff based on our interviews in the 2020-2021 and 2024-2025 timeframe efforts, for any organizations that are looking to increase their cybersecurity capabilities:

- MTSC-ISAC Membership plus attend annual MTS conference
- Cybersecurity Maturity Model Certification (CMMC) Level 2.0 compliance at each University
- Updated Researcher Cybersecurity CMMC checklist for DoD Projects
- Detailed SOP between each ship and it's University IT shop
- Develop ship specific Cyber Security Response Plans and test them
- Conduct tabletop ICS cybersecurity events for both ships and WHOI IS
- North American Treaty Organization (NATO) Science and Technology Organization (STO) Centre for Maritime Research and Experimentation (CMRE). This is a world-class scientific research and experimentation facility located in La Spezia, Italy. It is an executive body of the NATO Science and Technology Organization (STO) and focuses on the maritime domain to address the defense and security needs of the Alliance. CMRE conducts scientific research and technology development in various areas, including ocean science, modeling and simulation, acoustics, and more. The facility operates two research vessels, the NATO Research Vessel Alliance and the Coastal Research Vessel Leonardo, which are used for conducting NATO research and supporting commercial or government organizations within NATO nations.

Peregrine is already a member of these organizations, and if desired, we can brief these organizations on a regular basis on any critical issues, as the MTSC-ISAC promotes and facilitates maritime cybersecurity information sharing, awareness, training, and collaboration efforts between private and public sector stakeholders. If desired, our team can also monitor and review all of the updates in requirements/mandates that would need to abide by from wrt a cyber incident, as shown in Section 3.4. Peregrine will also support the upgrade to Safety Management System (SMM) section 7.10 - regarding Cyber Risk Management which will need to be updated / revised for inclusion in our manual. Revision / update of our SMM policy, US Coast Guard Vessel Security Plans, Cyber Security Incident Plans or other shipboard procedures to adhere to current UNOLS / NSF standards.

For the 2024-2025 effort, the Peregrine SMEs in conjunction with the University staff developed the specific documents to ensure from a Safety Management System (SMS) requirements, that they had a section for Cyber Risk Management (CRM) Instructions to employ heightened awareness and to establish procedures to respond to this situation on the vessels that they operate. This is a required document, from a safety aspect and will need to be updated in 2025 to match the new ABS Cyber Survey and USCG cybersecurity regulations.

Moving forward on the Cyber Security (CS) CS-1 notation, the Peregrine SMEs finalized documentation submissions (Subsection 1/7) are to be provided for surveys of notation requirements detailed in Section 2/Tables 3 and 4, respectively. As these initial surveys were conducted in October 2024 before the final draft of the new US Coast Guard cybersecurity mandate, our SMEs went on board to confirm that the following are documented and implementation-verified moving forward:

- Vessel cybersecurity risk management system (CRMS) (Appendix 3)
- Processes and programs detailed in Section 2/Tables 3 and 4, as applicable
- Maintenance procedures supporting Company-required maintenance of OT systems, IT systems, and CRMS protections as detailed on original equipment manufacturer (OEM solution provider documentation
- Vessel-specific cybersecurity programs and capabilities including Functional Design Document description (Appendix 2) and CS-1 Requirements listed in Section 2/Table 3
- The Peregrine SMEs icw WHOI are also determining which cyber risk management systems that have been surveyed to the satisfaction of the attending Surveyor to the full requirements of this Guide as applicable, where approved by the Committee, may be classed and distinguished in the ABS Record by the notation CS-1.

Here are the specific CS-1 Submittals that we complied with for this latest cyber assessment:

- Company and Vessel Cybersecurity Representative or Organization (1 August 2023) Company to provide documentation identifying person(s) responsible for cybersecurity, the organizational location within the Company, and quality or cybersecurity certificates (e.g., ISO/IEC 27001, ISA/IEC 62443, or ISASecure® certifications).
- Company and Vessel Cybersecurity Policies and Procedures (1 August 2023) Company to provide documentation detailing cybersecurity policies and implementation procedures applied to employees, suppliers, and contractors.
- Company and Vessel Cybersecurity Incident Response (IR) and Recovery. Documentation describing Company and vessel cybersecurity incident response team(s), including response and recovery responsibilities, capabilities, and staffing.
- Company and Vessel OT and Related IT System Architecture. Descriptions and diagrams detailing the OT and connected IT systems architecture that are suitable for performing a physical and cybersecurity risk assessment.
- Company and Vessel OT and Related IT Risk Assessment and Management Plan. A cybersecurity risk assessment of the OT and connected IT systems architecture, and a risk management plan derived from that assessment.
- Company and Vessel Cybersecurity Risk Management System (CRMS) Design. Operating and Maintenance Procedures CRMS functional description document (FDD) that includes a functional architecture with an inventory and descriptions of cybersecurity protective equipment; logical and procedural protections (see Appendix 3); operating and maintenance procedures; cybersecurity event/incident corrective/ preventive action procedures; auditing and test procedures; and, change management procedures.
- Company and Vessel Cybersecurity Training Program. Documentation detailing cybersecurity training records and training content concerning cyber hygiene and support of specialized cybersecurity functions.
- Company and Vessel OT/IT Management of Change (MOC) Procedures. Documentation detailing change management and configuration control policies and procedures applied during vessel operation, including OT, OT-connected IT, and CRMS software and computer hardware registries

From the new USCG Cybersecurity regulation, there are three main parts, specifically to maintain a Cybersecurity Plan, to designate a Cybersecurity Officer (CySO) and to implement a Cyber Incident Response Plan as part of the annual audit. Here are specific requirements:

- Cybersecurity Plan (101.630) with seven account security measures, plus four security measure requirements as well as two data security measure requirements
- Designate a Cybersecurity Officer to implement the Cybersecurity Plan and the Cyber Incident Response Plan, plus lead the annual audit as well as cybersecurity inspections. The CYSO must be accessible to the USCG 24/7.
- Drills and exercises (101.635) – at least once a year

- Records and Documentation (101.640) – for at least 2 years
- Specific Cybersecurity Measures (101.650) – sections a through I, covering all security measures, training, risk management, supply chain and physical security
- It also references the need for a Cyber Incident Response Plan

Peregrine as part of its future support has already started to draft the Cybersecurity Plan using this format:

- Cybersecurity Organization / CySO
- Personnel Training
- Drills and Exercises
- Records and Documentation
- Communications
- Cybersecurity Measures/Routine Maintenace
- Cybersecurity Measures/Account Measures
- Cybersecurity Measures/Physical Security
- Cybersecurity Measures/Supply Chain
- Cybersecurity Plan/Audits
- Cybersecurity Measures: Risk Management (Reports)
- Cybersecurity Measures: Risk Management (Vulnerabilities)
- Cybersecurity Measures: Resilience
- Cybersecurity Measures: Risk Management (Assessment)

As noted, the Peregrine SMEs have also drafted proposed cybersecurity playbooks for the following Cyber Incident Response Plan:

- Compromised Navigation System
- Malicious Software
- Loss of Internet
- Ransomware
- Phishing

Finally, we noted that the there are other entities that are striving to make strategic and appropriately aggressive decisions to incentivize industry adoption, improve productivity, and appropriately balance risks and opportunities. One area of expertise from Peregrine perspective that we can help WHOI to focus on more is cybersecurity for industrial control systems, and our SMEs can ensure that that these additional requirements are met moving forward to comply with these requirements:

- Maritime Cyber Risk Management in Safety Management Systems (MSC-FAL.1/Circ. 3)
- Resolution MSC.428(98) - Maritime Cyber Risk Management in Safety Mgmt Systems
- National Cybersecurity Center of Excellence - Securing Water and Wastewater Utilities Sector and Chemical Companies to outline cybersecurity posture at OT / ICS threats plus Responding /Recovering from cybersecurity incidents within the Manufacturing Sector
- NIST SP 800-160 - Revised guidance on engineering trustworthy secure systems
- Government Accountability Office (GAO) - Reports offshore oil and gas infrastructure faces cyber risks
- DHS - Testing environments to safeguard transportation infrastructure

5. Conclusion

Over the last decade, Peregrine has been collaborating from a mission-based aspect to lock down ICS using cyber risk assessments, plus we have developed a TTX that utilizes five playbooks of possible scenarios:

- Compromised Navigation System
- Malicious Software
- Loss of Internet
- Ransomware
- Phishing

We believe that taken together, this approach will facilitate the norming of best practices/processes and risk frameworks for cyber vulnerabilities on common OT platforms/components beyond the scope of traditional Information Security. Taken together, Peregrine has developed a holistic ICS /OT cybersecurity training regime

for the maritime environment that can be utilised in a number of unclassified as well as military / commercial environments, and we have the staff available to support any entities as needed.

Ethics Declaration: I did not need an ethical clearance for this research paper.

AI Declaration: I did not use an AI tool in the development of the paper.

References

International Maritime Organization (IMO) released a Maritime Safety Committee (MSC) resolution that addresses Maritime Cyber Risk Management in Safety Management Systems (MSC-FAL.1/Circ. 3), MSC-FAL.1-Circ.3-Rev.3.pdf

Cybersecurity Implementation for the Marine and Offshore Industries, ABS CyberSafety® Vol 2, Aug 23, Cybersecurity Implementation for the Marine and Offshore Industries

U.S. Coast Guard Issues Final Rule & Request for Comments on New Cybersecurity Regulations for the Marine Transportation System, 17 January 2025, Cyber Regulations Fact Sheet for Public Release_CLEAN_15JAN2025.pdf

The Guidelines on Cyber Security Onboard Ships, BIMCO, 4 January 2021, The Guidelines on Cyber Security Onboard Ships

Automated Threat Modeling using Artificial Intelligence on User Stories within the SDLC to Generate Security Tasks

Shantu Asif Hossain
SAP America Inc., Newtown Square, USA

Shantu_33_cse@yahoo.com

Abstract: This research presents an AI-driven system that integrates automated threat modeling directly into the Software Development Lifecycle (SDLC) during the early user story creation phase. Traditional threat modeling is often manual, delayed, and disconnected from developer workflows, resulting in missed vulnerabilities and reactive security measures. The proposed system employs a Large Language Model (LLM)-based Threat Modeling Engine to analyze user stories-textual descriptions of software features from an end-user perspective-and identify potential security threats. Leveraging advanced LLM algorithms, the system correlates detected risks with known threat patterns (e.g., STRIDE) and dynamically maps them to multiple pluggable security and compliance standards such as NIST CSF, ISO 27001, PCI DSS, HIPAA, SOC 2, OWASP, and GDPR. The engine automatically generates prioritized, technical security tasks aligned with these standards, which are seamlessly integrated into popular development tools like Jira, GitHub Issues, or Azure DevOps. This process enables proactive, traceable, and consistent enforcement of security controls throughout the development workflow, reducing human error and enhancing compliance with relevant regulations. A human-in-the-loop approval mechanism ensures full oversight and iterative refinement of threat modeling outputs. Furthermore, the system parses security standard documents in native formats (e.g., PDFs) to maintain up-to-date mappings without manual intervention. By embedding intelligent threat mitigation early in the SDLC, this research improves software security posture, development efficiency, and compliance adherence. It addresses a critical gap in current DevSecOps practices by automating and contextualizing security task generation from user stories, enabling development teams to build secure, compliant software aligned with national and international cybersecurity frameworks.

Keywords: Automated threat modeling, STRIDE, SDLC, Compliance automation, DevSecOps

1. Introduction

Software systems increasingly face threats that exploit gaps between design and implementation. Threat modeling is designed to systematically identify, classify, and mitigate potential security risks before they materialize, and it is typically applied during the early phases of the SDLC, such as requirements engineering and architectural design. Threat modeling is critical for anticipating vulnerabilities early, but practice often lags due to manual effort, timing, and weak integration with developer workflows. Recent work shows both the promise of LLMs for security tasks and the risks they introduce, motivating careful engineering of AI-assisted pipelines that are accurate, traceable, and compliant (Ferrag et al., 2024; Tete, 2024).

Large Language Models (LLMs) are increasingly applied to software engineering tasks such as code completion, vulnerability detection, and natural language requirements analysis. Their ability to parse user stories and generate structured outputs suggests an opportunity to embed automated threat modeling directly into developer workflows.

However, prior approaches have struggled with three interrelated challenges: (RQ1) effectively identifying security threats from user stories despite their informal and often ambiguous nature, (RQ2) aligning those threats with overlapping and frequently updated compliance standards, and (RQ3) ensuring that generated security tasks can be seamlessly integrated into development toolchains while preserving traceability and auditability. This research is structured around three questions, which are presented in Section 3.

Research contributions:

- **Threat modeling (RQ1):** An AI-driven engine derives STRIDE-based threats directly from user stories, producing actionable, traceable subtasks early in the SDLC.
- **Compliance alignment (RQ2):** Generated subtasks map consistently to major security and compliance frameworks (NIST CSF, ISO/IEC 27001, PCI DSS, HIPAA, SOC 2, OWASP ASVS, GDPR), ensuring control-level accuracy and adaptability as standards evolve.
- **Workflow integration (RQ3):** A system architecture processes user stories, maps them to standards, and outputs candidate tasks. Human reviewers validate these tasks before backlog entry, ensuring traceability, auditability, and readiness for agile tools like Jira.

2. Background and Related Work

2.1 Threat Modeling and STRIDE

Threat modeling provides a structured methodology for anticipating vulnerabilities. The STRIDE framework-Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privilege - remains widely used. However, STRIDE is under-utilized in early SDLC phases due to the high manual effort involved (Jedrzejewski et al., 2024).

2.2 Compliance Frameworks

Software systems must address overlapping frameworks including NIST CSF, ISO/IEC 27001, PCI DSS, HIPAA, SOC 2, OWASP ASVS, and GDPR. Mapping requirements across these frameworks is complex and error-prone (Kosenkov et al., 2025). Although official crosswalks exist (e.g., NIST SP 800-53 to ISO/IEC 27001), they remain largely static (Mussmann, 2021; NIST, 2023). Security Compass (2023) and others call for more dynamic mapping.

2.3 AI for DevSecOps

Recent surveys (Ferrag et al., 2024; Prates et al., 2025) highlight how LLMs are being integrated into vulnerability detection, policy extraction, and security automation. Koball et al. (2024) emphasize both their utility and new risks (e.g., prompt injection, data leakage). Work by Parvathinathan et al. (2025) and Yigit et al. (2025) demonstrates AI-powered continuous security, while UK DSIT (2024) highlights risks of LLM deployment.

2.4 User Stories and Automated Analysis

User stories are central to agile development. Dalpiaz (2025) and Casillo et al. (2022) showed that NLP can extract requirements quality and privacy needs. Robeer et al. (2016) extracted conceptual models from user stories, while Aboukadri et al. (2025) applied RAG-augmented LLMs to derive access-control policies. These works validate user stories as viable inputs for automated reasoning.

2.5 LLMs for Threat Modeling

Elsharef et al. (2024) investigated LLMs for facilitating threat identification. Jedrzejewski et al. (2025) studied STRIDE-based classification for network systems, and Nagaraja et al. (2025) modelled threats in LLM-based healthcare. Yang (2024) fine-tuned LLMs for banking, while Jedrzejewski, Fucci and Adamov (2025) introduced ThreMoLIA, a framework that produced structured threat outputs for large language model–integrated applications. These works demonstrate feasibility, but integration into developer workflows and compliance-aware task generation remains limited.

2.6 Research Gap

There is no unified approach that: (1) ingests user stories, (2) performs STRIDE analysis, (3) maps findings across multiple frameworks, and (4) emits developer-ready tasks with traceability inside issue trackers. This gap motivates our work.

3. Research Questions

3.1 RQ1 – Threat Modeling Capability

Can a large-language-model–powered engine derive meaningful STRIDE-based threats directly from user stories and express them as actionable, traceable security subtasks?

3.2 RQ2 – Compliance Alignment and Sustainability

Can the system automatically map detected threats to established security and compliance standards (e.g., NIST CSF, ISO 27001, PCI DSS, HIPAA, SOC 2, OWASP, GDPR) with measurable consistency*?*

3.3 RQ3 – Workflow Integration and Adoption

Can AI-generated security tasks from user stories be integrated into issue-tracking systems (e.g., Jira, GitHub Issues, Azure DevOps) while maintaining traceability and compliance alignment?

4. Methodology

4.1 Dataset Construction

To ground the study in realistic development practices, user stories were exported directly from Jira in CSV format (see Appendix A). This export preserved Jira's native schema - including fields such as Issue key, Issue id, Status, Summary, Description, Issue Type, Priority etc. Maintaining this structure ensured that the dataset accurately reflected real developer workflows, avoided artificial formatting bias, and allowed generated outputs to be validated in the same environment used by practitioners.

4.2 Motivation and Rationale

User stories represent the earliest and most consistent artifacts of agile workflows, making them the most practical entry point for automated threat modeling. By maintaining Jira-like structure, the evaluation reflects how such a system would operate in practice without introducing artificial complexity.

- For RQ1, the dataset tested whether the engine could generate STRIDE-based threats directly from user stories.
- For RQ2, it allowed mapping the generated subtasks to compliance frameworks, validating regulatory relevance.
- For RQ3: Workflow integration was simulated by formatting system outputs in Jira-compatible JSON. Each generated subtask included its STRIDE category, threat description, risk assessment (E1–E3, D1–D4, N1–N2), security task title with acceptance criteria, and compliance references. This approach would allow integration into agile workflows.

4.2.1 Reference annotations

Table 1 presents the baseline system prompt used for STRIDE threat analysis. The prompt instructs the model to output a risk-ordered (High to Low) table of unique mitigation findings (security subtasks), including assets, short threat scenarios, qualitative risk factors (E1–E3, D1–D4, N1–N2), acceptance criteria, and mappings to NIST CSF, ISO 27001, PCI DSS, HIPAA, SOC 2, OWASP, and GDPR.

The design of this baseline was informed by the OWASP Threat Modeling Process (OWASP, (n.d.-b)) to ensure methodological rigor and reproducibility. This expert baseline was then used to benchmark all evaluated LLMs (see Section 5).

Table 1: System Prompt for Automated Threat Modeling using STRIDE

System Prompt
As an experienced security architect and certified threat modeling expert, do the STRIDE threat modeling to generate security sub tasks, each sub task is treated one finding from the threat modeling exercise. Generate as much security sub tasks as required uniquely as mitigation. Your output should have following table format and Order by RISK = High to Low:
- ID (e.g., 1,2,3 etc.)
- Issue_key (e.g., {US/Bug/RS}-{id})
- Security_Task_ID (e.g. 1-100)
- STRIDE Category (e.g., Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege)
- Assets (e.g., a descriptive name that clearly identifies the asset)
- Threat Description (explain shortly what the threat with a meaningful short scenario is description)
- Risk (e.g., High, Medium, or Low risk calculate based on E1-E3, D1-D4, N1-N2 mentioned below. LIST like E1-D4=Yes/No/NA, N1-N2=5)
- Security Task Title (e.g., Implement, Protect etc.)
- User Story Acceptance Criteria as Mitigation (bullet points)
- NIST CFS Reference (e.g., each reference should have title with its equivalent id e.g. PR.AA-01: Identities and credentials for authorized users, services, and hardware are managed by the organization).
- ISO 27001 (provide full title of the security reference as is in the reference documentation).
- PCI DSS Reference (provide full title of the security reference as is in the reference documentation).
- HIPAA (provide full title of the security reference as is in the reference documentation).
- SOC 2 (provide full title of the security reference as is in the reference documentation).
- OWASP (provide full title of the security reference as is in the reference documentation).
- GDPR (provide full title of the security reference as is in the reference documentation).

Calculate the RISK based on Qualitative Risk Model such as "Ease of exploitation" and "The impact". e.g., "Ease of exploitation""

- E1. Can an attacker exploit this remotely?

- E2. Does the attacker need to be authenticated?

- E3. Can the exploit be automated?

"The impact" calculated from "damage potential" and "number of components that are affected by a threat" e.g., "damage potential":

- D1. Can an attacker completely take over and manipulate the system?

- D2. Can an attacker gain administration access to the system?

- D3. Can an attacker crash the system?

- D4. Can the attacker obtain access to sensitive information such as secrets or PII?

"Number of components that are affected by a threat" e.g.,

- N1. How many connected data sources and systems can be impacted?

- N2. How many layers into infrastructure components can the threat agent traverse?

Now, perform the STRIDE Threat Modeling exercise for following user story:

{User_Story}

4.3 Demonstration Case: OWASP Juice Shop

To illustrate the methodology, we applied it to the OWASP Juice Shop, a deliberately insecure web application widely used for security training and research (OWASP (n.d.-a)). Its architecture and vulnerability landscape make it an ideal candidate for STRIDE-based analysis.

In threat modeling practice, a diagram is often drawn by a security expert to capture assets, actors, components, and trust boundaries. From a SDLC perspective, however, reference architectures are also frequently attached to user stories as part of design discussions. These diagrams are produced either by solution architects or as outputs of exploratory research tasks, serving as guidance for developers during implementation.

For this study, we used the official reference architecture of OWASP Juice Shop as a representative diagram (Figure 2). It illustrates the frontend (Angular + Material), backend (Node.js/Express with Sequelize and NoSQL), and supporting file system components (SQLite, content folders), including integration with external identity providers such as Google via OAuth 2.0.

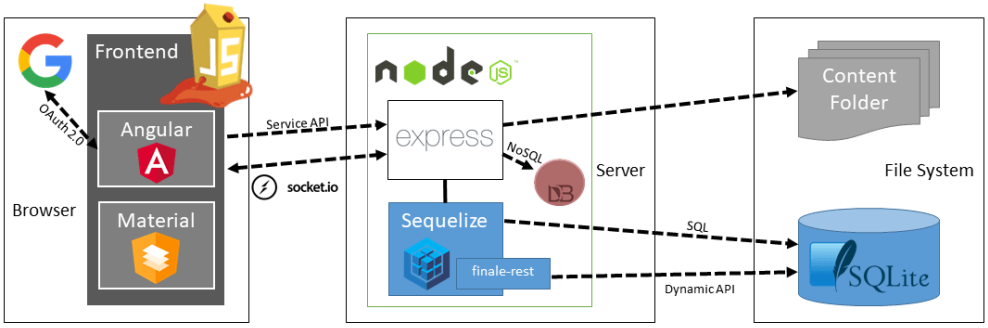


Figure 1: Reference architecture of the OWASP Juice Shop application OWASP (n.d.-c)

Following user story was generated using ChatGPT-5 model from https://chatgpt.com feature called “Add photos & files” by uploading Figure 1, and is used in this research:

> *“As a user, I want to log in with my Google account (OAuth 2.0) via the Angular frontend so that I can securely access application features served by the Node.js/Express backend, which stores data in SQLite/NoSQL and serves files from the content folder.”*

The AI-driven engine was tasked with producing STRIDE-based threats, mitigation subtasks, and compliance mappings. Results are presented in Section 5.

4.4 Models Evaluated

For this conceptual study, we evaluated outputs generated by two advanced large language models: ChatGPT-5 (OpenAI, via ChatGPT plus subscription: https://chatgpt.com) and Gemini-Pro (Google DeepMind, via AI Studio: https://aistudio.google.com/). Both represent cutting-edge generative AI systems optimized for reasoning, structured task generation, and security-relevant content creation.

4.4.1 GPT-5 (OpenAI)

GPT-5 was selected as the primary model for its reliable structured outputs (tabular, JSON, compliance-aligned), extended context window for processing detailed user stories with traceability, and practical availability through OpenAI's API and enterprise tools.

4.4.2 Gemini-Pro (Google DeepMind)

Gemini-Pro was used to test robustness and cross-model consistency. It generated STRIDE-based subtasks mapped to major frameworks (NIST CSF, ISO 27001, PCI DSS, HIPAA, SOC 2, OWASP, GDPR), delivered outputs in Markdown and JSON for reproducibility, and showed strength in cross-mapping ambiguous cases. Accessible via Google AI Studio, it supports research and academic validation.

4.5 System Architecture and Workflow Integration

To evaluate workflow integration (RQ3), we designed a system architecture that processes user stories and generates candidate security tasks (Figure 2).

The workflow begins with a user story submitted by a developer. The AI-driven threat modeling engine analyzes the story, references established security standards and generates candidate security tasks. These tasks flow into a human-in-the-loop review stage:

- If approved, tasks are added to the backlog with preserved traceability and compliance mappings.
- If rejected, they are iteratively refined by the engine until they meet acceptance criteria.

This architecture ensures outputs are technically sound, traceable, and standards-aligned, while remaining directly compatible with agile project management environments such as Jira, GitHub Issues, and Azure DevOps.

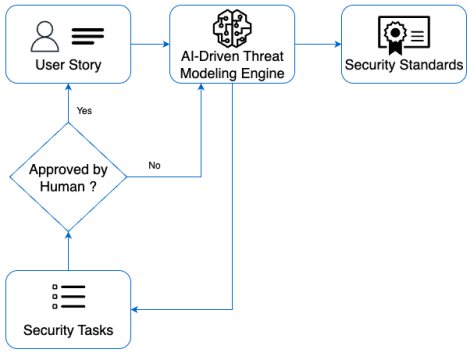


Figure 2: System architecture of the AI-driven threat modeling engine

The architecture consists of four main components:

- **User Story Input:** Represents developer-authored requirements serving as the entry point.
- **AI-Driven Threat Modeling Engine:** Automatically analyzes user stories, classifies threats via STRIDE, and maps results to compliance frameworks.
- **Security Standards:** Provide security standard references (e.g., NIST CSF, ISO 27001, PCI DSS etc.) to align subtasks with regulatory expectations when required, while also supporting ingestion of organization-specific policies in native formats (e.g., PDF) for LLM-based parsing and application.
- **Human-in-the-Loop Review:** Ensures governance by allowing experts to approve, refine, or reject tasks before integration into development backlogs.

4.6 Security Considerations needed in AI Integration

Integrating AI with development artifacts like user stories risks information disclosure, as project systems may contain sensitive logic, design details, or credentials. To mitigate this, the design incorporates safeguards:

- Deploying the LLM on-premise or in a private cloud, or integrating with a company-managed AI service under established governance controls, to prevent uncontrolled data exposure.
- Applying personal and sensitive data anonymization to mask identifiers such as names, email addresses, access tokens, or system details before processing.
- Using encryption and audit logging within the Jira plugin to secure records of all interactions.
- Conducting compliance checks aligned with GDPR and SOC 2 principles to ensure lawful data processing.

Prior research (Ferrag et al., 2024; Tete, 2024) highlights privacy leakage as a key vulnerability. Embedding these safeguards ensures automation improves security without creating new exposure risks.

5. Results

5.1 RQ1 - Threat Modeling Capability

The OWASP Juice Shop login story was evaluated using ChatGPT-5 (14 subtasks, IDs 1–14, see Figure 3) and Gemini-Pro (10 subtasks, IDs 101–110). Both models generated structured threats across all six STRIDE categories, demonstrating that LLMs can approximate expert-level threat modeling.

ID	Issue Key	Security Task ID	STRIDE Category	Assets	Threat Description	Risk (E1–E3, D1–D4, N1–N2)	Security Task Title	User Story Acceptance Criteria (Mitigation)
1	US-001	1	Spoofing	OAuth 2.0 Authorization Code Flow	Attacker intercepts/replays OAuth codes/tokens to impersonate the user during Google login.	High (E1=Yes, E2=No, E3=Yes; D1=Yes, D2=Yes, D3=No, D4=Yes; N1=5, N2=4)	Enforce PKCE, State & TLS	Enforce TLS 1.2+ with HSTS; Require PKCE; Validate state/nonce; Short-lived & rotating tokens
2	US-001	2	Tampering	Service API (Node.js/Express)	Client manipulates request parameters/IDs to bypass authorization (BOLA/BFLA).	High (E1=Yes, E2=No, E3=Yes; D1=Yes, D2=Yes, D3=Yes, D4=Yes; N1=5, N2=4)	Object-Level Authorization (BOLA)	Server-side ownership checks; Default deny RBAC/ABAC; Validate JWT scope; Strict schema validation
3	US-001	3	Elevation of Privilege	Sequelize ORM + SQL/NoSQL Query Layer	Injection via unsafe queries grants admin-level operations.	High (E1=Yes, E2=No, E3=Yes; D1=Yes, D2=Yes, D3=No, D4=Yes; N1=4, N2=3)	Eliminate Injection & Enforce Least Privilege	Parameterized queries only; Disable/review raw queries; Least-privilege DB roles; Automated SAST/DAST
4	US-001	4	Information Disclosure	Token Storage (Browser/App)	ID/Access tokens stored in local/session storage leaked via XSS or logs.	High (E1=Yes, E2=No, E3=Yes; D1=No, D2=No, D3=No, D4=Yes; N1=4, N2=3)	Harden Token Storage	HttpOnly/Secure/SameSite cookies; Don't log tokens; Rotate & revoke tokens
5	US-001	5	Spoofing	OAuth Redirect URI & Client Registration	Manipulated redirect_uri or rogue client_id causes token leakage to attacker.	High (E1=Yes, E2=No, E3=Yes; D1=Yes, D2=No, D3=No, D4=Yes; N1=4, N2=3)	Strict OAuth Client & Redirect URI Policy	Register exact URIs; Enforce HTTPS; Validate iss/aud/azp claims
6	US-001	6	Tampering	JWT Verification Pipeline	Attacker tampers with alg/claims or uses unsigned tokens.	High (E1=Yes, E2=No, E3=Yes; D1=Yes, D2=Yes, D3=No, D4=Yes; N1=4, N2=3)	Robust JWT Verification	Verify signature; Enforce alg allow-list; Check exp/nbf/aud/iss; Reject "none"
7	US-001	7	Information Disclosure	Content Folder (Static Files)	Path traversal or direct URL access exposes sensitive files.	High (E1=Yes, E2=No, E3=Yes; D1=No, D2=No, D3=No, D4=Yes; N1=4,	Secure File Serving	Serve via API only; Allow-list paths; Enforce RBAC

Figure 3: STRIDE-based subtasks generated for OWASP Juice Shop user story by ChatGPT-5 (partial view)

5.1.1 Overlaps

Overlapping findings captured the essential STRIDE threats. Each ID (e.g., ChatGPT 1) refers to the corresponding Security_Task_ID field in the model outputs, full output files can be found in Appendix.

- **Spoofing:** OAuth replay and state validation (Gemini 101; ChatGPT 1, 5).
- **Tampering:** Token theft and unsafe API modification (Gemini 104, 108; ChatGPT 2, 6).
- **Elevation of Privilege:** Backend authorization weaknesses (Gemini 102; ChatGPT 3).
- **Information Disclosure:** Token storage and database leakage (Gemini 103, 109; ChatGPT 4, 7, 10, 12).
- **Denial of Service:** Flooding and database exhaustion (Gemini 106, 110; ChatGPT 8, 14).
- **Repudiation:** Missing audit logs (Gemini 105; ChatGPT 9).

5.1.2 Unique findings

Distinct outputs from each model highlight complementary strengths, expanding overall coverage beyond the overlaps:

- **ChatGPT-5:** CSRF protection (11), secrets management (12), session fixation (13), CORS misconfiguration (10).
- **Gemini-Pro:** CSP/HSTS (107), malicious file uploads (108), environment/config security (109), database query overload (110).

Six subtasks overlapped directly, while each model contributed four to five unique findings. This yields about 20 distinct subtasks rather than 24. ChatGPT-5 provided granular developer-focused mitigations, whereas Gemini-Pro emphasized compliance and operational safeguards. Combined, the two models produced a balanced and comprehensive set of security tasks aligned with the STRIDE methodology.

5.2 RQ2 - Compliance Alignment and Sustainability

The OWASP Juice Shop login story evaluation generated 14 subtasks from ChatGPT-5 (IDs 1–14, see Figure 4) and 10 subtasks from Gemini-Pro (IDs 101–110). Each subtask was mapped to STRIDE categories, OWASP Top 10 categories, and major compliance frameworks (NIST CSF, ISO 27001, PCI DSS, HIPAA, SOC 2, OWASP ASVS, GDPR). Together, the two models demonstrated strong alignment with security and compliance standards, though some coverage gaps remain.

ID	NIST CSF Reference	ISO 27001 Reference	PCI DSS Reference	HIPAA Reference	SOC 2 Reference	OWASP Reference	GDPR Reference
1	PR.AA-01: Identities and credentials for authorized users, services, and hardware are managed	A.9.2.4 Management of secret authentication information of users	Req. 8: Identify users and authenticate access to system components	§164.312(d) Person or entity authentication	CC6.2 Logical access security measures	ASVS V2.1 Authentication design; OWASP API Security: API2 Broken Authentication	Art. 32 Security of processing
2	PR.AA-04: Access to data, assets, and devices is limited to authorized users, processes, and devices	A.9.1.2 Access to networks and services controlled	Req. 7: Restrict access to cardholder data by business need to know	§164.312(a)(1) Access control	CC6.4 Logical access restricted	OWASP API Security: API1 Broken Object Level Authorization	Art. 25 Data protection by design and by default
3	PR.DS-06: Mechanisms for data integrity are implemented and tested	A.12.6.1 Technical vulnerability management; A.9.2.3 Privileged access rights	Req. 6.2/6.3 Secure software and applications; Req. 6.5 Common coding vulnerabilities	§164.312(c)(1) Integrity	CC6.6 Controls protect against injection	OWASP Top 10 A03:2021 Injection	Art. 32 Security of processing
4	PR.DS-05: Access permissions and authorizations are managed, incorporating least privilege and separation of duties	A.10.1.1 Cryptographic controls; A.9.4.1 Information access restriction	Req. 3.4 Protect stored account data; Req. 8.3 Strong authentication	§164.312(e)(1) Transmission security; §164.312(a)(2)(iv) Encryption	CC6.1 Policies and procedures constrain logical access	ASVS V3.4 Sensitive data storage; Token Binding Cheat Sheet	Art. 32 Integrity and confidentiality
5	PR.AA-01: Identities and credentials are managed	A.9.2.1 User registration and de-registration	Req. 8.2 Secure protocols for authentication	§164.312(d) Authentication	CC6.2 Logical access security measures	ASVS V2.2 OAuth/OIDC verification	Art. 32 Security of processing
6	PR.DS-06: Integrity checking mechanisms	A.14.1.2 Securing application services on public networks	Req. 6.5.9 Verify authenticity of sessions/tokens	§164.312(c)(1) Integrity	CC7.2 Monitoring and anomaly detection	ASVS V3.2 Token integrity; JWT Cheat Sheet	Art. 32 Integrity
7	PR.DS-05: Access permissions managed	A.9.4.1 Information access restriction	Req. 7.2 Protect public access to sensitive resources	§164.312(a)(1) Access control	CC6.4 Restrict logical access	OWASP Top 10 A01:2021 Broken Access Control; A05:2021 Security Misconfig	Art. 5(1)(c) Data minimization; Art. 32 Confidentiality

Figure 4: Compliance Alignment Mappings by ChatGPT-5 (partial view)

5.2.1 OWASP category coverage

Both models consistently addressed high-priority OWASP risks, cross-checked against the known vulnerability categories of the OWASP Juice Shop application (OWASP, n.d.-d) for better traceability, while each contributed unique coverage areas:

- Broken Authentication: OAuth/token replay, redirect-URI validation, and session management (ChatGPT 1, 5, 13; Gemini 101, 102).
- Broken Access Control (BOLA): Object-level authorization and file/path controls (ChatGPT 2, 7; Gemini 104).
- Injection: Unsafe ORM and query handling (ChatGPT 3; Gemini 103).
- Sensitive Data Exposure / XSS: Token leakage, client storage, CSP/HSTS, and environment protections (ChatGPT 4; Gemini 107, 109).
- Security Misconfiguration: Static file serving, JWT/CORS defaults, and config protection (ChatGPT 7, 10; Gemini 109).
- Denial of Service / Resource Controls: Flooding, upload quotas, and DB query overload (ChatGPT 8, 14; Gemini 106, 110).
- Logging & Monitoring / Repudiation: Audit logging and log retention (ChatGPT 9; Gemini 105).
- CSRF / Session Management: Explicit mitigations surfaced by ChatGPT (10, 11).
- File Handling Security: Malicious upload checks introduced by Gemini (108).

5.2.2 Gaps

Neither model generated explicit coverage for Insecure Deserialization, Vulnerable/Outdated Components (dependency or SBOM management), or XML External Entities (XXE).

5.2.3 Compliance framework alignment

- **NIST CSF (USA):** The generated subtasks mapped directly to key NIST Cybersecurity Framework categories—PR.AC (Access Control), PR.DS (Data Security), DE.CM (Detection and Monitoring), PR.IP (Protective Technology), and PR.PT (Protective Measures). By translating abstract framework principles into actionable user-story tasks, this research addresses the long-standing gap between high-level policy frameworks and day-to-day developer adoption. It effectively operationalizes "secure-by-design" practices at the earliest stage of the SDLC, ensuring security is built in rather than bolted on. Such integration is directly aligned with U.S. national cybersecurity priorities, including supply-chain protection and small-to-medium business adoption, thereby reinforcing the broader national interest in embedding NIST CSF as a baseline across industries.
- **PCI DSS:** Tasks addressing token protection, session handling, and audit logging map to PCI DSS requirements for data protection, authentication, and monitoring.
- **GDPR (EU):** Encryption, pseudonymization, logging, and retention subtasks provide safeguards consistent with GDPR Article 32.
- **ISO 27001 / HIPAA / SOC 2:** Gemini-Pro in particular emphasized alignment with ISO control domains (A.9, A.12, A.14, A.17) and HIPAA/SOC 2 obligations, strengthening governance aspects.

Table 2: Summary Table

OWASP Category	ChatGPT-5 Security Task IDs	Gemini-Pro Security Task IDs	Compliance Frameworks
Broken Authentication	1, 5, 13	101, 102	NIST CSF PR.AC; PCI DSS 8; ISO 27001 A.9
Broken Access Control (BOLA)	2, 7	104	NIST CSF PR.AC; SOC 2 CC6
Injection	3	103	PCI DSS 6.5; OWASP ASVS 5; ISO 27001 A.14
Sensitive Data Exposure / XSS	4	107, 109	GDPR Art.32; HIPAA §164.312; NIST CSF PR.DS
Security Misconfiguration	7, 10	109	NIST CSF PR.IP; ISO 27001 A.12; SOC 2 CC7
Denial of Service / Resource Limits	8, 14	106, 110	NIST CSF DE.CM; ISO 27001 A.17; SOC 2 CC7
Logging & Monitoring / Repudiation	9	105	NIST CSF DE.AE; PCI DSS 10; ISO 27001 A.12
CSRF / Session Management / CORS	10, 11	-	OWASP ASVS 4.2; NIST CSF PR.AC
File Handling Security	-	108	ISO 27001 A.12; NIST CSF PR.IP

The combined results show that ChatGPT-5 emphasized developer-focused mitigations (CSRF, session fixation, secrets management, CORS) while Gemini-Pro emphasized compliance-linked and operational tasks (CSP/HSTS, file upload risks, environment configs, DB query overload). Together, they produced stronger compliance alignment than either model individually. While some OWASP categories remain uncovered, the ability of both LLMs to map user-story–derived subtasks to STRIDE, OWASP, and compliance frameworks demonstrates a meaningful step toward automated compliance-aware threat modeling.

5.3 RQ3- Workflow Integration and Adoption

The system is designed for direct adoption into DevSecOps workflows by producing machine-readable outputs in a Jira-compatible JSON format. Each generated security task is represented as a discrete JSON object that links to the originating user story and includes identifiers, STRIDE category, asset, threat description, qualitative risk dimensions, mitigation tasks, acceptance criteria, and compliance mappings. This enables seamless import into

platforms such as Jira, GitHub Issues, or Azure DevOps, where security tasks can be managed alongside functional requirements.

Both models, ChatGPT-5 and Gemini-Pro, generate JSON outputs with minor format differences. ChatGPT-5 represents risk as a structured object with subfields (E1–E3, D1–D4, N1–N2) and uses human-readable keys (e.g., "Threat Description"), while Gemini-Pro encodes risk as a compact string and employs snake_case keys (e.g., "Threat_Description", "STRIDE_Category"). Despite these schema variations, both preserve the same essential content: traceability from threat to mitigation and alignment with compliance frameworks (NIST CSF, ISO 27001, PCI DSS, HIPAA, SOC 2, OWASP, GDPR)

Figure 5 illustrates an example export (Security_Task_ID 3 from ChatGPT-5), where an injection threat is captured as a JSON object with structured acceptance criteria and mapped standards. Equivalent artifacts from Gemini-Pro confirm that the format is reproducible across models, reinforcing interoperability.


```
{
  "ID": 3,
  "Issue_key": "US-001",
  "Security_Task_ID": 3,
  "STRIDE Category": "Elevation of Privilege",
  "Assets": "Sequelize ORM + SQL/NoSQL Query Layer",
  "Threat Description": "Injection via unsafe queries grants admin-level operations.",
  "Risk": {
    "Overall": "High",
    "E1": "Yes",
    "E2": "No",
    "E3": "Yes",
    "D1": "Yes",
    "D2": "Yes",
    "D3": "No",
    "D4": "Yes",
    "N1": 4,
    "N2": 3
  },
  "Security Task Title": "Eliminate Injection & Enforce Least Privilege",
  "User Story Acceptance Criteria as Mitigation": [
    "Use parameterized queries/ORM bind parameters only",
    "Disable raw queries or review/approve tightly",
    "DB accounts least-privilege with separated roles",
    "Automated SAST/DAST for injection patterns"
  ],
  "NIST CSF Reference": "PR.DS-06: Mechanisms for data integrity are implemented and tested.",
  "ISO 27001": "A.12.6.1 Technical vulnerability management; A.9.2.3 Management of privileged access rights",
  "PCI DSS Reference": "Req. 6.2/6.3: Secure software and applications; Req. 6.5: Address common coding vulnerabilities",
  "HIPAA": "\u00a7164.312(c)(1) Integrity",
  "SOC 2": "CC6.6 Controls protect against injection and other attacks",
  "OWASP": "OWASP Top 10 A03:2021 Injection",
  "GDPR": "Art. 32 Security of processing"
},
```


Figure 5: AI output in JSON format for security Task ID 3 generated by ChatGPT-5.

Key fields enabling workflow integration include:

- Issue Key / User Story Linkage – ensures context with development tasks.
- Security Task ID – unique identifier for tracking.
- STRIDE Category & Threat Description – structured threat modeling reference.
- Risk (E1–E3, D1–D4, N1–N2) – explicit qualitative assessment.
- Security Task Title & Acceptance Criteria – actionable, testable controls.
- Compliance References – mapping to standards for audit readiness.

This export format provides a standards-annotated, automation-ready schema that ensures traceability and reduces manual transcription. Human review remains the final governance step, but the JSON representation enables scalable integration of AI-generated security tasks into enterprise pipelines.

6. Discussion

The demonstration shows that an LLM-driven threat modeling engine can be used to extract STRIDE-relevant threats directly from plain-language user stories and translate them into actionable, standards-mapped security subtasks. In the OWASP Juice Shop login example the LLM Models produced several distinct subtasks covering all STRIDE categories; each subtask included an affected asset, a concise threat description (written so non-experts can understand why it matters), a qualitative risk profile, acceptance criteria, and mapped controls. The JSON export format preserved traceability to the parent story and supplied the fields required for programmatic ingestion into issue trackers.

Strengths of the approach include: (1) shifting threat analysis left into the requirements phase, enabling earlier remediation decisions; (2) producing developer-friendly, actionable subtasks (titles + acceptance criteria) that reduce ambiguity; and (3) coupling threats with framework references (NIST CSF, ISO 27001, PCI DSS, HIPAA, SOC 2, OWASP, GDPR) to support compliance-aware engineering. The human-in-the-loop review remains essential for governance - to refine mappings, reject false positives, and adapt outputs to project context.

Important challenges surfaced during evaluation: certain STRIDE classes (notably repudiation and denial-of-service) are context-dependent and harder for a text-only model to detect reliably; exact control-ID fidelity sometimes requires manual refinement (parser/mapping improvements); and the demonstration was a

simulated integration (JSON export) rather than a live plugin, so operational behaviors such as CI/CD integration, roles, permissions, and end-to-end latency remain to be validated at scale. Overall, the results support feasibility and practical value while highlighting concrete engineering improvements needed for production use.

7. Limitations and Future Work

This work was constrained by its reliance on Jira-formatted user stories, a single demonstration case, and general-purpose LLMs that were not security-tuned. SME input was limited, and integration was shown only through JSON exports rather than production-grade CI/CD testing.

Future efforts should broaden datasets across industries and user-story types (e.g., complex stories with attachments, bug reports, research tasks), fine-tune LLM models, and expand SME validation. Additional directions include handling non-text artefacts such as UML diagrams and embedding the system into live CI/CD pipelines to assess practical scalability.

8. Conclusion

This paper presents a lightweight, LLM-powered approach to automated threat modeling that operates at the user-story stage of the SDLC. The engine converts plain-language user stories into STRIDE-classified threats, actionable security subtasks with acceptance criteria, and mapped compliance references; outputs are emitted as machine-readable, Jira-compatible JSON to preserve traceability and enable programmatic ingestion. The OWASP Juice Shop demonstration produced a comprehensive set of subtasks that aligned with known vulnerability categories and illustrated practical workflow compatibility. While further work is needed to improve coverage (some threat types and third-party/component risks) and to validate live operational integration, the approach shows promise for embedding standards-aware security earlier in development, reducing rework, and improving the auditability of security activities. All artifacts for reproducibility (datasets, annotations, JSON exports) are available in the project repository (see Appendix A).

Ethics Declaration: This research did not involve human participants, personal data collection, or sensitive experiments; therefore, formal ethical clearance was not required.

AI Declaration: AI tools were used extensively in this work to improve readability, grammar, and overall structure. Their role was limited to editorial support, and they did not contribute to the design of the study, data analysis, or interpretation of results. All substantive content, findings, and conclusions are entirely the author's own, and all AI-generated content was carefully reviewed and validated by the author.

References

Aboukadri, S. et al. (2025) 'Leveraging RAG and LLMs for access-control policy extraction from user stories in Agile software development', *IEEE Access*. Available at: https://www.researchgate.net/publication/393424984 (Accessed: 20 September 2025).

Casillo, F. et al. (2022) 'Detecting privacy requirements from user stories with NLP', *Information and Software Technology*. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0950584922000246 (Accessed: 20 September 2025).

Dalpiaz, F. (2025) 'Automated analysis of user story requirements', in *Natural Language Processing for Requirements Engineering*. Springer. Available at: https://link.springer.com/chapter/10.1007/978-3-031-73143-3_12 (Accessed: 20 September 2025).

Elsharef, I. et al. (2024) 'Facilitating threat modeling by leveraging large language models', *AISCC@NDSS 2024*. Available at: https://www.ndss-symposium.org/wp-content/uploads/aiscc2024-16-paper.pdf (Accessed: 20 September 2025).

Ferrag, M.A., Maglaras, L., Shu, L. and Derhab, A. (2024) 'Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities', *arXiv preprint* arXiv:2405.12750. Available at: https://arxiv.org/abs/2405.12750 (Accessed: 20 September 2025).

Jedrzejewski, F.V. et al. (2024) 'Threat modeling of ML-intensive systems', *ACM Proceedings*. Available at: https://dl.acm.org/doi/10.1145/3644815.3644975 (Accessed: 20 September 2025).

Jedrzejewski, F.V. et al. (2025) 'LLMs' suitability for network security: A STRIDE-based case study', *arXiv preprint*arXiv:2505.04101. Available at: https://arxiv.org/abs/2505.04101 (Accessed: 20 September 2025).

Jedrzejewski, F.V., Fucci, D. and Adamov, O. (2025) 'ThreMoLIA: Threat Modeling of Large Language Model-Integrated Applications', arXiv preprint arXiv:2504.18369. Available at: https://arxiv.org/abs/2504.18369 (Accessed: 21 September 2025).

Koball, C. et al. (2024) 'Machine learning security: Threat model, attacks, and mitigations', *IEEE Computer*. Available at: https://dl.acm.org/doi/abs/10.1109/MC.2024.3396357 (Accessed: 20 September 2025).

Kosenkov, O. et al. (2025) 'Requirements engineering for regulatory compliance: A systematic mapping study', *Information and Software Technology*. Available at: https://www.sciencedirect.com/science/article/pii/S0950584924002271 (Accessed: 20 September 2025).

Mussmann, A. (2021) 'Mapping the state of security standards mappings', *GITO Verlag White Paper*. Available at: https://library.gito.de/wp-content/uploads/2021/08/L4_Mussmann-Mapping_the_State_of_Security_Standards_Mappings-305_c.pdf (Accessed: 20 September 2025).

Nagaraja, N. et al. (2025) 'Cyber threat modeling of an LLM-based healthcare system', *ICISSP/SCITEPRESS*. Available at: https://www.scitepress.org/Papers/2025/132897/132897.pdf (Accessed: 20 September 2025).

NIST (2023) 'Mapping SP 800-53 Rev. 5 to ISO/IEC 27001:2022', *NIST CSRC*. Available at: https://csrc.nist.gov/projects/olir/informative-reference-catalog/details?referenceId=155#/ (Accessed: 20 September 2025).

OWASP (n.d.-a) OWASP Juice Shop Project. Available at: https://owasp.org/www-project-juice-shop/ (Accessed: 21 September 2025).

OWASP (n.d.-b) Threat Modeling Process. Available at: https://owasp.org/www-community/Threat_Modeling_Process (Accessed: 21 September 2025).

OWASP (n.d.-c) OWASP Juice Shop Application Architecture. Available at: https://owasp.org/www-project-juice-shop/ (Accessed: 21 September 2025).

OWASP (n.d.-d) Vulnerability categories. Available at: https://help.owasp-juice.shop/part1/categories.html (Accessed: 21 September 2025).

Parvathinathan, K. et al. (2025) 'AI-powered DevSecOps for continuous security and compliance', *ResearchGate Preprint*. Available at: https://www.researchgate.net/publication/395004862 (Accessed: 20 September 2025).

Prates, L. et al. (2025) 'DevSecOps practices and tools: A multivocal literature review', *International Journal of Information Security*. Available at: https://link.springer.com/article/10.1007/s10207-024-00914-z (Accessed: 20 September 2025).

Robeer, M. et al. (2016) 'Automated extraction of conceptual models from user stories', *IEEE Requirements Engineering Conference*. Available at: https://webspace.science.uu.nl/~dalpi001/papers/robe-luca-werf-dalp-brin-16-re.pdf(Accessed: 20 September 2025).

Security Compass (2023) 'Mapping OWASP ASVS to ISO 27001', *Whitepaper*. Available at: https://www.securitycompass.com/whitepapers/mapping-security-requirements-to-standards-owasp-asvs-to-iso-27001/(Accessed: 20 September 2025).

Tete, S.B. (2024) 'Threat modelling and risk analysis for large language model (LLM)-powered applications', *arXiv preprint* arXiv:2406.11007. Available at: https://arxiv.org/abs/2406.11007 (Accessed: 20 September 2025).

UK DSIT (2024) 'Cyber security risks to artificial intelligence', *UK Government*. Available at: https://www.gov.uk/government/publications/research-on-the-cyber-security-of-ai/cyber-security-risks-to-artificial-intelligence (Accessed: 20 September 2025).

Yang, S., Wu, T., Liu, S., Nguyen, D., Jang, S. and Abuadbba, A. (2024) 'ThreatModeling-LLM: Automating threat modeling using large language models for banking system', arXiv preprint arXiv:2411.17058. Available at: https://arxiv.org/abs/2411.17058 (Accessed: 21 September 2025).

Yigit, Y. et al. (2025) 'Generative AI and LLMs for critical infrastructure protection', *Sensors*. Available at: https://www.mdpi.com/1424-8220/25/6/1666 (Accessed: 20 September 2025).

9. Appendix A. Dataset and Repository

All datasets, , figures, and outputs used in this study are available in the companion GitHub repository: 👉 https://github.com/ShantuApps/Reseach-Automated-Threat-Modeling-using-AI

SOC Puppets: How Whaley's *Theory of Outs* and 'Noisy' Sock Puppets Encouraging Discovery of Network Deception Could Enhance Security Operations Center Analysis

Tim Pappa[1] and Keyur Rajyaguru
[1]Analyst1, Reston, Virginia, USA

tim.pappa@analyst1.com
kpr@umd.edu

Abstract: This practitioners' position paper suggests an unorthodox approach integrating Whaley's *Theory of Outs* and "turnabout" deception techniques to encourage an attacker's discovery of deception on a network. Although the late American deception and communication researcher Barton Whaley appeared to differentiate the categorization of deception techniques such as "turnabout" and planning for backup deception techniques if discovered or if the deception appeared to fail, we explore the integration of these approaches to deception design in a cyber context, mainly in situations where analysts in the Security Operations Center (SOC) are looking for higher fidelity alerting on anomalous events and suspected attacker activity on the network. Because attackers appear to generally demonstrate greater confidence in their network movement after discovering what they believe is deception, we visualize how 'noisy' controlled deception sock puppets inside of a network prompting optimized query returns on their content could draw attackers to later stage deception functions and effects, and more enhanced SOC analysis following alerting on those functions and effects. This practitioners' position paper suggests our unorthodox approach offers an alternative strategic and tactical approach to collaborative cyber deception design and SOC alerting, by highlighting 'noisy' deception on a network to lure and influence attackers.

Keywords: Cyber deception, Theory of outs, Turnabout, Security operations center, Cyber deception design

1. Introduction

This proposed integrated modeling is foundationally based on an increasingly practiced unorthodox operational approach to industry cyber deception, where deception is intentionally 'noisy' and staged to be discovered. Fundamentally, this idea contradicts the traditional SOC analysis approach where noisy detections are tuned out to prevent alert fatigue. However, our more recent experience in industry cyber deception operations has involved more of an organizational approach to deception functions and effects, where generally there is less direct engagement with attackers than we previously experienced when supporting law enforcement operations online. Because there is generally less direct engagement with attackers, there is often less observed attacker behaviors. Designing deception functions and effects without a baseline of behavior is arguably difficult. There must be some application of behavioral and cognitive frameworks in the absence of demonstrated attacker behaviors.

Much of the research in the past several years related to behavioral and cognitive frameworks has concentrated on cognitive vulnerabilities and attacker behaviors on a host or network when influenced by a range of cognitive biases. This research concentration on cognitive biases reflects growing efforts to explore a cognitive theory of deception, instead of what has historically been a focus on physical deception cues. Many of these same researchers concentrating on the effect of cognitive biases have encouraged a more "adaptive" cyber deception design and build, primarily based on deceptive signaling targeting attackers designed to influence changes in their attacker behaviors (Cranford et al., 2020; Cranford et al., 2021). Gonzalez et al. (2020) advocated for more dynamic forms of defense, with deception integrated into engineering design. Gonzalez et al. wrote that designing effective defense must consider the knowledge of human behaviors, namely the way people make decisions and explore a network. Incorporating these kinds of "dynamic forms of defense" in deception could result in a collection of uniquely rich attacker behaviors that demonstrate an individual attacker's reconnaissance and decision-making. That behaviorally rich collection could inform more behaviorally adaptive cyber deception operations.

Aggarwal et al. (2024) wanted to build on their findings from a prior study where participants behaving like attackers appeared to identify cognitive biases, exploring whether attackers showed any preference for targeting specific areas of a network and if attackers revealed any behavioral patterns when operating in those specific areas. Their findings suggested attackers did demonstrate cognitive biases, such as sunk cost fallacy and default effect. The goal of this follow-up research was to determine if defenders could create more dynamic

responses to attackers based on linking identified cognitive biases with "behavioral patterns" observed in similar network environments. Some of these behavioral patterns appeared to reflect default effect, where attackers are more likely to choose a "preset alternative" or default when making decisions. This finding is like naturalistic decision-making research, where attacker decision makers appear to process information differently, based on their experiences making similar decisions when presented with similar information in the past (Du et al., 2023; Yuill, Denning, and Feer, 2007). Aggarwal et al. in their recent study found that attackers appeared to demonstrate this default effect most commonly in the reconnaissance phase of their attack cycle. Aggarwal et al. (2024) found other cognitive biases as well, such as availability bias or when people appear to make choices based on the information that comes to mind first, for example. Their findings in this study highlighted what they characterized as the strength of the "presence of a bias", based on whether participants behaving like attackers found a task or function on a network to be interesting or not or to what degree they found it interesting. Aggarwal et al. suggested that there are other functions that attackers may find less interesting, that may exploit other cognitive biases because of the kind of processing someone might engage in when completing some task or function considered to be mundane or expected.

Shinde and Doshi (2024) presented an approach to modeling and exploiting the cognitive biases of attackers when planning for "active deception", highlighting the cognitive biases fundamental attribution error and confirmation bias. Shinde and Doshi concentrated on human attackers, arguing that sophisticated attacks are still largely driven by human attackers, and human attackers are vulnerable to these cognitive biases. Their approach noted that even cyber deception largely developed with artificial intelligence is still based on a belief that attackers will be rational, when most attackers are not necessarily rational but vulnerable to these biases and other decision-making influences. Shinde and Doshi largely focused on the information gathering process of attackers and how they might overattribute factors in an environment to a human defender rather than assessing an environment they observe and interact with based on network and host functions. An example might be an attacker who believes a Bloodhound attack pathway is only vulnerable because defenders have nestled that pathway with decoys that alert on attackers, but perhaps in this example that attack pathway is vulnerable because defenders have overlooked vulnerabilities on that pathway because they defaulted to automated monitoring and mitigation management.

Aggarwal et al. (2025) tested approximately a half dozen cognitive biases with participants in similar studies, finding similar results like the cognitive biases studies above. But another finding was that participants' experience and knowledge did appear to limit the potential effect of attempts to induce cognitive biases in participants. While those more experienced participants may have been less affected by these tasks, they also appeared to be more willing to take risks during scenarios in this study. The findings suggested there may be a shift in risk preferences among experienced attackers that do not reflect the commonly reported findings on risk aversion.

In contrast, the information security industry's typical approach to cyber deception has largely been static and functional, however. There is a modestly growing market for commercial off-the-shelf software for deception and packaged honeypots. Those honey- prefix builds have evolved significantly in the past twenty years, but these 'turnkey' applications are generally limited and instrumental to a function on a network, rather than a function behaviorally responsive to observed attackers (Javadpour et al., 2024). These approaches are largely mitigation techniques. Lansborough et al. (2021) noted that defensive network systems are historically statically configured, generally failing to adapt to an attacker or adapt in ways that an attacker predicts or easily counters.

Strand's industry handbook on cyber deception, "The Art of Active Defense", represented almost a decade of advocating for more offensive or proactive responses to attackers. Strand (2017) wrote that network defenders should engage attackers whenever possible, essentially 'annoying' them by wasting their time and forcing them to make additional moves, which are likely to increase collection on an attacker toward some attribution. Strand often characterized cyber deception as "annoyance". Strand's handbook included a considerable number of techniques that can be applied to attackers, but these engagements would usually occur during some phase of incident response or threat hunting. These techniques were generally reactive to attackers, designed to frustrate attackers. There was limited knowledge in those moments of an attacker's behaviors or capabilities. This approach largely continues to be the standard, except when augmented with increasingly automated and curated deception software. There may be growing recognition that simply suggesting there is deception on a network is effective. There are still few organizations that appear willing to communicate that (Ferguson-Walter, 2018, 2019, 2024). We agree with this research approach into the cognitive vulnerabilities of attackers, but we also suggest that there are deception techniques or approaches that have still rarely been applied or operationalized by network defenders or cyber deception practitioners. These techniques or approaches have

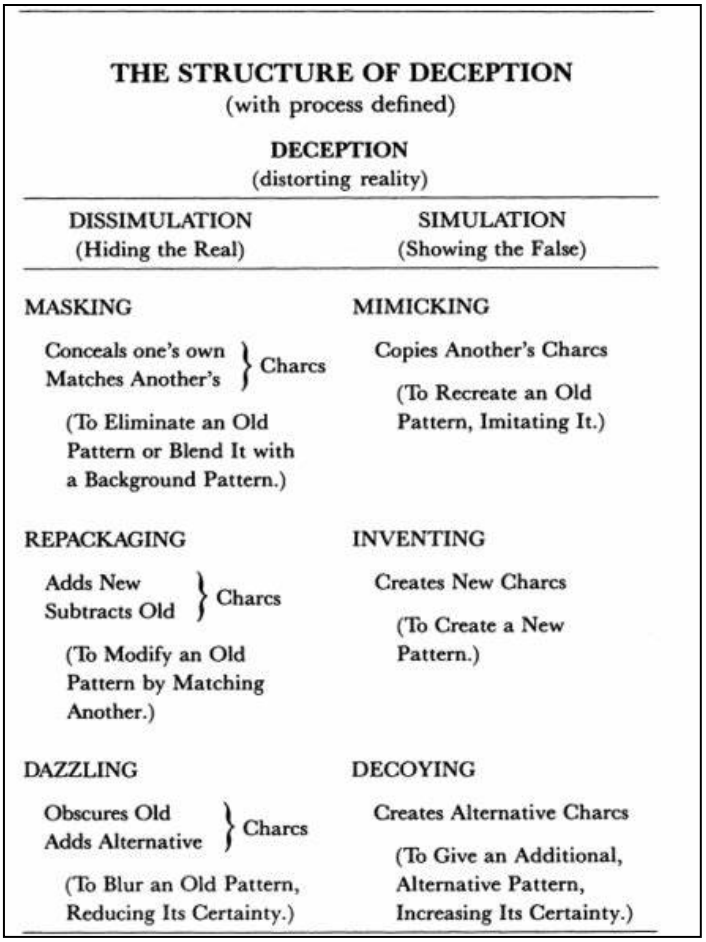
been discussed in historical military deception research and commentary, but there has been little to no known application in cyber deception and SOC contexts.

This paper is organized as follows. First, this paper will briefly discuss some of Whaley's historical work on deception. His shared framework is the foundation of the deception techniques discussed in this paper, including his *Theory of Outs* and his discussion of the "turnabout" techniques. Second, this paper will include a working characterization of the integrated approach of this theory and these lesser-known techniques in a cyber context. Third, this position paper will highlight the function of SOC across industry and organization and how Deception operations integrate within the defensive methodologies. Finally, this paper will visualize the application of this integrated approach to cyber deception to help enhance the SOC's analysis of suspected and known attacker events and incidents when incorporating 'noisy' sock puppets.

2. Related Work on Whaley's Characterizations of Deception and Misperception

The Bell-Whaley (1982, 2017) framework is the starting point for the author's cyber deception design process. This framework has been discussed extensively in literature, primarily in military deception literature (Daniel and Herbig, 1982; Martin, 2008; Whaley, 1982; Lloyd, 2003; Smith, 1992). Bell and Whaley structured this framework to demonstrate that these techniques are simultaneously *dissimulating* and *simulating*. Dissimulating in any deception scenario might include masking, or hiding the real by making it invisible, for example. Simulation might include inventing or showing the false by fabricating something.

Whaley (1974) emphasized the simultaneous dynamic of this deception framework in an early characterization where he described a scenario of burying a bag of gold coins in his backyard. He wrote that if he is burying or hiding (*dissimulating*) the golden coins to hide his wealth, he is simultaneously demonstrating or suggesting (*simulating*) that the golden coins are not at his home or that the golden coins are located somewhere else, and he is also simulating that he is not wealthy but in fact poor.



THE STRUCTURE OF DECEPTION
(with process defined)

DECEPTION
(distorting reality)

DISSIMULATION (Hiding the Real)	SIMULATION (Showing the False)
MASKING	**MIMICKING**
Conceals one's own / Matches Another's } Charcs	Copies Another's Charcs
(To Eliminate an Old Pattern or Blend It with a Background Pattern.)	(To Recreate an Old Pattern, Imitating It.)
REPACKAGING	**INVENTING**
Adds New / Subtracts Old } Charcs	Creates New Charcs
(To Modify an Old Pattern by Matching Another.)	(To Create a New Pattern.)
DAZZLING	**DECOYING**
Obscures Old / Adds Alternative } Charcs	Creates Alternative Charcs
(To Blur an Old Pattern, Reducing Its Certainty.)	(To Give an Additional, Alternative Pattern, Increasing Its Certainty.)

Figure 1: The Bell-Whaley deception framework

Whaley (1980) wrote that he believed there was a "poverty of theory" about misperception of simulation and dissimulation as a foundational deception model, meaning people may generally understand surprise or how they feel to be surprised but they may not recognize the psychological underpinnings of surprise.

Whaley described misperception as both *qualitative* and *quantitative*. The qualitative depth of misperception is variety, namely that there are several ways in which someone or a system can be deceived. Someone or a system could misperceive something that is observed or communicated. Whaley wrote that the quantitative depth or dimension of misperception is intensity, namely the degree or even the frequency to which misperception occurs. Whaley (1984) added that when we consider how we characterize misperception in terms of the variety of cues, in that perception and the intensity of or degree of that misperception, we could think of players on a baseball team as a characterization of strength in the context of misperception.

Whaley wrote that while we can count each player on a team in terms of numbers to represent strength, we also can and should consider how each player has qualitatively differentiated roles and positions on a team. Whaley noted that each of those players individually may be better or worse than each other and may be better or worse than other players on other teams. Whaley wrote that "we tend to think we know who the players are", but that can change depending on which baseball teams are playing against each other and how those players on each team are performing and how they chose to perform.

3. A Working Characterization of Integrated "Turnabout" and "Outs" Deception Planning

The purpose of this paper is to explore how someone like an attacker might respond behaviorally and cognitively to deception, especially when an attacker believes they have discovered deception. Whaley characterized "turnabout" as a context where deception appears to have been discovered, and there is an opportunity to respond with deception. "Turnabout" was a historical and practiced technique in military deception, but Whaley found that planning for "turnabout" in contemporary deception operations in the military appeared to be considerably rare in contrast to the practice of "turnabout" in historical military examples.

Whaley shared an example of "turnabout" in a plot by Soviet intelligence to add laxatives to the saltshakers in the Radio Free Europe cafeteria in Prague. The Soviet asset directed to add the laxatives confessed to security personnel and then American intelligence the plot before any saltshakers were compromised, but American intelligence allowed this plot to be revealed publicly to expose this attempt by Soviet intelligence to frighten Radio Free Europe personnel and locals working for this American government news media organization. American intelligence allowed this plot to proceed for a while before exposing it.

Whaley wrote about what has been called a "Let's Double Back trick" or a "double back" as another representation of "turnabout". This generally means someone has 'doubled back' to a starting point in an ongoing deception, whether that deception was planned or not. The audience or target of that deception may presume that someone in some pattern is following a sequence of activity that will lead to an anticipated outcome, rather than returning to a starting point. The "double back" is surprising, then. But deception planners planned for it.

Whaley referred to a deception operation in the 1970s where the Central Intelligence Agency had been collecting imagery of the Soviet Union from their KH-11 "Keyhole" satellite, searching for potential SALT agreement violations related to missile production. Because the Soviets were aware of this collection, they would often relocate or hide missile infrastructure. When the CIA later incorporated an unknown new design feature in the KH-11 satellite, where collection was transferred to another satellite and then transmitted to the ground rather than directly from the KH-11 to ground, the Soviets could no longer detect any data imagery transmissions from the KH-11 to ground. The Soviets' conclusion appeared to be that the satellite was "dead".

The Soviets stopped hiding their missiles during these flyovers because they believed the satellite was not collecting imagery. But the CIA continued collection and transmission, until the Soviets later discovered from an asset that the CIA had modified the satellite.

Whaley referred to another example of "double back" from an operation known as MOONSHINE from World War II. The British had begun distorting German radar detection by using older models of radar emulation technology. When the British began 'doubling back' or reintroducing older versions of their radar technology, the Germans struggled to accurately detect enemy aircraft. The number or scale of approaching British aircraft was distorted because the Germans had configured their detection systems to look for more advanced British radar hardware. The Germans eventually discovered this "double back" and adjusted their detection efforts. About a year after that discovery, the British reintroduced this early version of their radar emulation of an older

model again, which again distorted German detection efforts because they expected to see more advanced versions.

Whaley differentiated his *Theory of Outs* as planning for "alternative goals" for a deception if that deception is discovered or if the deception appears to have failed. Whaley wrote that deception rarely fails because there is generally some effect because of deception where the target is influenced in some manner, but deception planners in his experience consistently did not include options in their deception operations for accomplishing the deception or influence goal in some other way if the deception was discovered.

There are situations he wrote where a strategic goal of deception may appear to have failed, but a tactical goal of the same deception succeeds.

Whaley wrote about the "Delayed Message Trick" where the true operational plans are transmitted on a channel the target is monitoring, but the transmission is slowed or delayed in a manner where the target will be unable to react effectively. Whaley wrote that this approach is rare and should be used sparingly, but this is an example of planning for an "out" by managing the transmission of the operational plans but perhaps keeping those plans vague and making it appear there are technical issues with the target's monitoring infrastructure. There is still flexibility in this deception design to hold the target's attention and influence them, but to control the transmission or communication of that deception content.

Whaley referred to failed airborne parachute drops in Normandy in 1944 as an example of what appeared to be failed deception, but the deception was still effective. The goal of Allied airborne drops in Normandy was to slow or block the responding German reinforcements until the beachhead was secured. There were also "diversionary drops" of fake or dummy paratroop units. Whaley wrote that the Germans were immediately aware of the scattered airborne drops, but they assumed those scattered drops were deliberate rather than a mistake.

The Germans missed an opportunity to destroy or ignore those scattered and lightly armed parachute units and proceed directly to the beachhead, but instead because the Germans believed it was deliberate they allocated reinforcements to engage each of those scattered units slowly and cautiously. Whaley wrote that the unplanned failed airborne drops and scattered units appeared to have a better effect than the planned version because it still achieved the objective of slowing the Germans' response. Whaley emphasized the need for deception planners incorporating "outs" in their deception operation to consider how the target might misperceive deception it discovers or believes it has discovered.

This paper characterizes a working approach to deception planning or deception design integrating "turnabout" and this theory of "outs", where planners include the discovery of the deception into their design.

Whaley may have intended these approaches or techniques in deception to be separate given the context or situation, but the author suggests that in a naturalistic offline and online network attack environment there should be an integration of these techniques and layered deception design.

4. Role of the Security Operations Center (SOC) in Cyber Defense and how "Turnabout" Deception Strengthens Defensive Mechanisms

A Cyber Security Operations Center (SOC) is a unit within an organization responsible for monitoring and responding to cyber threats in real time. Generally, a team of cybersecurity analysts operates 24/7 to investigate alerts, determine their severity, and take necessary actions.

Deception operations can be integrated into alert pipelines. Some of the most common deception artifacts include false credentials stored in documents, or fake confidential documents, for example. Any alert from the honeypots enhances SOC operations by providing context about the attacker or insider threat behavior, objectives and methods leveraged. A fake but strategically placed honeytoken when accessed gives the SOC insights into what data types are targeted by the attackers.

Learning attacker techniques is a much-needed part of proactive and reactive cyber defense. These techniques become more useful when they are observed in the infrastructure the SOC is trying to protect. Since these activities are performed in an environment where we have full control, it becomes an attractive investment to put in some time and extract intelligence compared to the generic intelligence available about various attacker groups. deception alerts are inherently different from other SOC alerts. Generally, SOC alerts focus on anomalies in the network and user behavior to catch them. These anomalies can be in different forms like time of activity, length of engagement, nature of user performing actions, number of attempts made to access artifacts, and so

on. In deception alerts, those honeytokens are placed strategically where no one would attempt to access for their daily operations, only the ones with "extra curious" mindset or someone with intention of gaining unauthorized access to the confidential details would access it. The alerts become even more helpful when the tokens are placed further into the campaign from the entry point of the lure. Many times, people will stop progressing if they sense that they are not supposed to access this information. An attacker would not stop because that is exactly what they are after. A noisy deception environment would make an attacker more interested in advancing the attack even if they sense deception because they believe they can laterally move, and they have a strong foothold in the infrastructure. Some attackers may keep advancing just because of a great reconnaissance opportunity as they want to collect intelligence about the victim to use it later in other attack campaigns. This is where SOC can leverage "turnabout" deception techniques to study attacker behavior. These learnings can be funneled back to the detection logic and can be mapped elsewhere in the infrastructure. Feedback from SOC to a cyber deception team will also help to create more honey content that works. There are always some traces that attackers leave behind, even behaviors that bypass traditional rules or signature-based attack methodologies. Mapping those Indicators of Behavior within the enterprise network would improve security posture. SOC can leverage the use of MITRE ATT&CK framework to demonstrate how deception helps catalog attacker Techniques, Tactics, and Procedures (TTPs) before they can target valuable systems. These insights allow SOC to modify existing detections.

5. Visualizing a 'Noisy' Sock Puppet Deception Approach Integrating "Turnabout" and "Outs" Modeling

This cyber deception practitioner's position paper visualizes a scenario in which a controlled sock puppet account that appears to be demonstrably affiliated with a company has been consistently posting content on a social media platform about his or her infrastructure development work. This sock puppet by design has been making mostly generalized references to his or her projects, but there are also keywords in the content in blogs and posts that may be of interest to attackers who are also interested in targeting this company. This might include references to unique tools or projects that appear to be developmental, and involving access to company data considered to be sensitive or valuable if an attacker stole that data to sell it on a market.

Even if the attacker in this scenario obtains unauthorized access to this company's network simply by obtaining basic user account credentials, there will likely be some effort inside the network to verify whether the sock puppet account is a company employee or not and if that employee is posting or sharing content inside the network about similar projects. That kind of consistency in content creation and sharing over time by the same controlled sock puppet account is what makes that deception storyline plausible.

When the attacker in this scenario begins to query certain terms or keywords associated with this sock puppets projects and the area of interest to the attacker, there should be optimization of this sock puppet's content because of the consistent and considerable content creation internally by the sock puppet. This is the 'noisy' sock puppet content. While these are internal platforms, there are similar approaches to optimizing search content. Because there has been consistent content sharing by this sock puppet, there may have been prior false positives from other company employees curious about these projects or employees questioning the content he or she is posting. This is what makes this sock puppet content 'noisy'.

Network defenders in this scenario may have also placed simple decoys along this possible attack pathway for an attacker, so that the attacker will avoid what appear to be oddly placed honey files with likely alerting and perhaps obvious naming conventions on the files, such as "sensitive data_confidential_Q1". The author recognizes that this kind of naming convention may still appeal to attackers who will try to access this kind of honey file because of the naming convention or description of the data in the file in the title, however, this kind of approach would still represent the author's proposed integrated modeling of "turnabout" and "outs" deception planning. Network defenders want the attacker in this scenario to believe he or she has discovered these honey files in a plausible attack pathway, so that they will either trip those decoys and return with other credentials in the future or they will avoid these honey files and continue their movement, perhaps confident they have detected this deception.

In this scenario, once the attacker finds content of interest from his or her queries of keywords or terms of interest and the search return includes the sock puppet's content, the attacker may scrutinize this content less because he or she is familiar with this deception storyline, but unaware this is deceptive. The indexed content may include a hyperlink to another site on the network that suggests access to the data the attacker is seeking. When the attacker clicks on this link, he or she is transferred to a site where he or she can enter his or her stolen

user credentials. That access attempt will be successful, but there will be alerts. Then the attacker will be transferred to another controlled portal with further warning on that site regarding confidential access. When the attacker attempts to sign into this portal with the same credentials, the SOC will be alerted immediately. This incremental sequence of deception effects and function limits the number of false positives, discouraging employees from attempting to explore and access these sites and content. This sequence of deception functions also allows the SOC to observe a user's intent and to capture a high-fidelity alert.

6. Discussion

This visualized scenario demonstrates how a 'noisy' sock puppet's content creation inside and outside a network can not only gain the attention of an attacker and influence that attacker's cognitive and behavioral decision making throughout his or her attack cycle but mitigate and manage the behaviors of false positives or other employees who may also become interested in this 'noisy' content.

The visualized scenario applied the design framework proposed in this position paper. The attacker was presented with simple deception decoys along plausible attack pathways so that he or she would find that deception easily or avoid that deception and continue moving. We believed this might increase the attacker's confidence and suggested that the company's attempt at deception had been discovered and had failed.

Because the attacker may believe he or she has discovered this failed deception, he or she may scrutinize the remainder of files and data sites on the network. The sock puppet's content creation outside the network on a social media platform may have been surprising, but once the attacker quickly determines that this sock puppet appears to be a real employee at this company, the attacker again may scrutinize the sock puppet's content less. The attacker's belief that he or she has easily discovered attempts at deception and that he or she has successfully found a pathway to the data of interest associated with the sock puppet could have surprised or violated the attacker's expectations, but the author argues that the true violation of the attacker's expectations will occur when the attacker is transferred to controlled infrastructure and alerted on each of his or her attempts. The attacker will be quickly mitigated or will exit the network promptly when he or she realizes that the controlled portal and sock puppet content was also a deception, but much more material and influential.

Even in this relatively brief interaction and deception effect and function between an attacker and this controlled 'noisy' sock puppet content and deception, network defenders have demonstrably surprised or violated the attacker's expectations of network defense and deception capability. We argue that this is not only a reputational deterrence, but a strategic and tactical change in this company's interactions with this attacker and likely future attackers. This kind of cyber deception design modeling and application could enhance the SOC's confidence in their collection and analysis of these high-fidelity alerts based on 'noisy' sock puppets.

Ethics Declaration: Ethics clearance was not required for this paper.

AI Declaration: There was no use of AI in the development or at any other stage of this paper.

References

Aggarwal, P., Venkatesan, S., Youzwak, J., Chadha, R. and Gonzalez, C., 2024. Discovering Cognitive Biases in Cyber Attackers' Network Exploitation Activities: A Case Study. In *Human factors in cybersecurity. AHFE (2024) International conference*.

Aggarwal, P., Rubaiyet Nowmi, S., Du, Y. and Gonzalez, C., 2024. Evidence of Cognitive Biases in Cyber Attackers from An Empirical Study.

Aggarwal, A., Ferreria, Maria J., Aggarwal, P., Rajivan, P., and Gonzalez, C., 2025. Cognitive Biases in Cyber Attacker Decision Making: Translating Behavioral Insights into Cybersecurity. In 10th IEEE European Symposium on Security & Privacy Workshops, 4th Active Defense & Deception Workshop, Proceedings.

Bell, J.B. and Whaley, B., 2017. *Cheating and deception*. Routledge.

Bell, J.B. and Whaley, B., 1982. *Cheating: deception in war & magic, games & sports, sex & religion, business & con games, politics & espionage, art & science*. St Martin's Press.

Cranford, E.A., Gonzalez, C., Aggarwal, P., Tambe, M., Cooney, S. and Lebiere, C., 2021. Towards a cognitive theory of cyber deception. *Cognitive Science*, *45*(7), p.e13013.

Cranford, E., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M. and Lebiere, C., 2020. Adaptive cyber deception: Cognitively informed signaling for cyber defense.

Daniel, D.C. and Herbig, K.L., 1982. Propositions on military deception. *The Journal of Strategic Studies*, *5*(1), pp.155-177.

Du, Y., Prébot, B., Xi, X. and Gonzalez, C., 2023, January. A Cyber-War Between Bots: Human-Like Attackers are More Challenging for Defenders than Deterministic Attackers. In *HICSS* (pp. 856-865).

Ferguson-Walter, K., Shade, T., Rogers, A., Trumbo, M.C.S., Nauer, K.S., Divis, K.M., Jones, A., Combs, A. and Abbott, R.G., 2018. *The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception* (No. SAND2018-5870C). Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Ferguson-Walter, K., Shade, T.B., Rogers, A.V., Niedbala, E., Trumbo, M., Nauer, K., Divis, K., Jones, A., Combs, A. and Abbott, R., 2019. *Appendix to the Tularosa study: an experimental design and implementation to quantify the effectiveness of cyber deception* [online]

Ferguson-Walter, K.J., 2024. An empirical assessment of the effectiveness of deception for cyber defense.

Gonzalez, C., Aggarwal, P., Cranford, E.A. and Lebiere, C., 1825, January. Design of dynamic and personalized deception: A research framework and new insights for cyberdefense. In *Proceedings of the 53rd hawaii international conference on system sciences* (Vol. 1834).

Javadpour, A., Ja'fari, F., Taleb, T., Shojafar, M. and Benzaïd, C., 2024. A comprehensive survey on cyber deception techniques to improve honeypot performance. *Computers & Security*, p.103792.

Kambow, N. and Passi, L.K., 2014. Honeypots: The need of network security. *International Journal of Computer Science and Information Technologies*, *5*(5), pp.6098-6101.

Landsborough, J., Carpenter, L., Coronado, B., Fugate, S., Ferguson-Walter, K. and Van Bruggen, D., 2021, January. Towards Self-Adaptive Cyber Deception for Defense. In *HICSS* (pp. 1-10).

Lloyd, M., 2003. *The art of military deception*. Pen and Sword.

Martin, C.L., 2008. *Military deception reconsidered* (Doctoral dissertation, Monterey, California. Naval Postgraduate School).

Mokube, I. and Adams, M., 2007, March. Honeypots: concepts, approaches, and challenges. In *Proceedings of the 45th annual southeast regional conference* (pp. 321-326).

Shinde, A. and Doshi, P., 2024, May. Modeling cognitive biases in decision-theoretic planning for active cyber deception. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems* (pp. 1718-1726).

Smith, D.V., 1992. Military Deception and Operational Art.

Strand, J., Asadoorian, P., Robish, E. and Donnelly, B., 2013. *Offensive Countermeasures: The Art of Active Defense*. CreateSpace Independent Publishing Platform.

Whaley, B., 1982. Toward a general theory of deception. *The Journal of Strategic Studies*, *5*(1), pp.178-192.

Whaley, B., 1980. A Typology of Misperception or The Ways We Can Be Wrong. *Unpublished manuscript draft*.

Whaley, B., 1974. Deception: Its Decline and Revival in International Conflict. *Unpublished manuscript draft*.

Whaley, B., 2016. *Turnabout and Deception: Crafting the Double-cross and the Theory of Outs*. Naval Institute Press.

Yuill, J., Denning, D. and Feer, F., 2007, January. Psychological vulnerabilities to deception, for use in computer security. In *DoD Cyber Crime Conference* (Vol. 2007).

Work in Progress Papers

Exploring NIS2 Compliance in the Energy Sector Using AI-Driven Cyber Threat Intelligence

Jani Siivola, Rami Paronen, Uzair Tariq, Quyet Pham, Warren Villegas, Ilkka Tikanmäki[1] and Jyri Rajamäki[2]
Laurea University of Applied Sciences, Espoo, Finland

Ilkka.tikanmaki@laurea.fi

Abstract: The NIS2 Directive introduces stricter requirements for how essential entities, including energy-sector operators, must manage cybersecurity risks and report incidents. In practice, many organisations face difficulties in transforming these legal obligations into concrete, daily security operations, especially in operational technology (OT) environments where visibility, logging, and coordinated responses are often limited. This paper examines how SecureAI, an AI-based anomaly detection and enrichment tool within the Cyber Threat Intelligence (CTI) ecosystem, can help energy operators meet key NIS2 obligations. The study is based on a qualitative desk-research approach, a comparative mapping of SecureAI capabilities against NIS2 Articles 20-26, and a realistic OT case scenario based on recent intrusion patterns. Prior research shows that AI can detect industrial anomalies faster and more accurately than rule-based systems, and that automated CTI processing can turn raw alerts into structured and shareable intelligence. At the same time, NIS2 requires accountable use of such tools, meaning that human oversight, transparency of analysis, and reliable evidence generation must be part of AI-supported workflows. These requirements guided the assessment. The analysis shows that SecureAI supports several key NIS2-related tasks. It identifies unusual behaviour in network and host telemetry, enriches findings with asset information and event relationships, and produces structured alert objects that support operator decision-making. The CTI Framework then converts these enriched alerts into STIX/TAXII objects suitable for reporting, documentation, and intelligence exchange. The case scenario–an unauthorised remote-access intrusion followed by suspicious HMI-PLC activity–demonstrates how SecureAI can highlight the anomaly, provide context for understanding its impact, and supply material for reporting and further investigation.

Keywords: NIS2 Compliance, Critical energy infrastructure, AI-driven anomaly Detection, Cyber threat intelligence automation, Operational cyber defence

1. Introduction

The NIS2 Directive places stricter cybersecurity obligations on essential-sector operators, including organisations in the European energy sector (European Parliament and Council, 2022). These entities must implement risk-management measures and meet rigorous incident-reporting requirements, yet many struggle to translate NIS2's high-level legal expectations into daily operational practices. The challenge is particularly pronounced in OT environments, where limited visibility, sparse logging and legacy systems hinder effective threat detection and response.

Recent advances in artificial intelligence (AI) and cyber threat intelligence (CTI) offer promising ways to address these operational gaps. AI-driven anomaly-detection systems have demonstrated stronger performance than traditional rule-based tools (Govea et al., 2024), while automated CTI platforms can transform raw alerts into enriched, contextualised intelligence suitable for reporting and information sharing. These capabilities align with NIS2 requirements for early detection, structured incident handling and coordinated information exchange.

This study examines how SecureAI—an anomaly detection and alert enrichment tool—can support selected NIS2 operational obligations for energy sector OT environments. We map relevant NIS2 requirements to SecureAI and assess their application through a representative OT intrusion scenario. Integrating SecureAI into an energy sector OT environment will improve the operator's capability to fulfil NIS2 requirements for threat detection, incident handling, reporting, and information sharing.

The Research Question of this study is: Can the combination of AI-based anomaly detection and CTI automation provide an operational capability that meets the NIS2 requirements in energy sector OT environments?

2. NIS2 and the Energy Sector

Energy infrastructure is increasingly exposed to unprecedented cybersecurity threats that jeopardise the stability and reliability of critical systems (Ajayi, Alozie & Abioba, 2025). The NIS2 Directive (Directive (EU)

[1] https://orcid.org/0000- 0001-8950-5221

[2] https://orcid.org/0000–0003–4798–2462

2022/2555) strengthens cybersecurity requirements for essential entities, including electricity, gas, oil and hydrogen operators (European Parliament and Council, 2022). Several provisions relate directly to technical operations. Article 20 assigns management responsibility for approving cybersecurity measures and ensuring staff training. Article 21 outlines operational security measures, including monitoring, incident handling, backup and recovery, and vulnerability management. Article 23 introduces strict reporting timelines (24-hour early warning, 72-hour notification and one-month final report). Article 24 allows Member States to require the use of certified ICT products for selected security functions. Article 29 mandates structured information sharing among trusted entities (Ruohonen, 2024).

NIS2 also places greater emphasis on the resilience of OT environments, which are especially challenging due to legacy equipment, limited telemetry, and the need to maintain operational continuity. Unlike traditional IT systems, OT assets cannot be easily patched or rebooted without affecting safety or uptime, which complicates the implementation of NIS2's risk-management measures. Energy networks are also highly interconnected, meaning disruptions in one part of the system can cascade across regions or even countries. These operational realities increase the difficulty of implementing consistent monitoring, rapid incident response, and evidence-based reporting.

Recitals 51–52 encourage the use of automated and AI-based security tools, provided they operate transparently and support interoperability. This establishes a foundation for evaluating how SecureAI might support NIS2-aligned operational tasks in energy-sector OT environments. The combination of strict regulatory obligations and structurally challenging operational contexts creates a clear motivation for exploring whether AI-driven detection and CTI workflows can provide practical, scalable support for compliance.

3. Literature Review

AI-based anomaly detection has demonstrated notable benefits in industrial systems, outperforming rule-based approaches in detecting irregular or malicious activity under noisy OT conditions (Aydın, 2025; Govea et al., 2024). Machine-learning and deep-learning models can capture subtle behavioural deviations that would be invisible to threshold-based or signature-based detection methods. Studies also highlight that OT traffic often exhibits stable behavioural patterns, which makes it well-suited for modelling normal baselines—an advantage leveraged by anomaly-detection systems.

Research on compliance automation shows how AI can support regulatory obligations by processing requirements, generating structured documentation and reducing manual reporting workloads (Alevizos, 2025; Seethala, 2025). Automation can also reduce human error and improve consistency in the documentation submitted to regulators. NIS2's tight reporting deadlines make this especially relevant, as organisations often struggle with assembling reliable evidence within 24–72 hours.

Cyber threat intelligence (CTI) frameworks contribute to coordinated incident handling by correlating multi-source telemetry, enriching alerts, and sharing structured intelligence (Gong & Lee, 2021). Established platforms such as MISP, OpenCTI and the TAXII protocol ecosystem demonstrate the maturity of automated intelligence exchange. Research also suggests that integrated CTI workflows enable faster triage and reduce analyst workload during multi-stage attacks. Together, these strands suggest that AI-driven detection and CTI automation can meaningfully support the operationalisation of NIS2 in energy-sector OT environments, especially where resource constraints or telemetry gaps hinder traditional approaches.

4. Methodology

A qualitative desk-research approach was used to analyse how SecureAI can support operational tasks related to the NIS2 Directive in energy sector OT environments. Operational cybersecurity obligations - governance oversight, cybersecurity risk management, incident reporting, certification-related documentation and information sharing - were extracted from Articles 20, 21, 23, 24 and 29 of NIS2. As NIS2 combines organisational and technical requirements, the analysis focuses on obligations that technical systems can realistically support. Governance, accountability and certification remain organisational responsibilities. Auditability and traceability were included as interpretive criteria when assessing technical outputs. Each selected operational requirement was compared against SecureAI and CTI capabilities to determine where they support NIS2-aligned workflows. The resulting alignment is summarised in Section 5.

5. Results

SecureAI contributes most directly to Articles 21, 23 and 29 while providing supporting evidence for Articles 20 and 24. To illustrate how SecureAI supports NIS2-aligned workflows, we apply it to a representative OT intrusion scenario inspired by recent attacks reported in the energy sector (Dragos Inc., 2025; ENISA, 2024). An attacker exploits an exposed remote-access interface to enter a substation network and issues unauthorised HMI commands to alter PLC setpoints. Such sequences reflect real-world intrusions targeting substation protection schemes, where even minor configuration changes can result in significant electrical or safety impacts.

In a typical substation architecture, remote-access servers sit between IT and OT zones and are intended only for maintenance use. If security controls are weak—such as default passwords, unpatched VPN appliances or inadequate segmentation—these access points become attractive attack vectors. Once inside, adversaries can move laterally across engineering workstations, historians, or HMIs to identify PLCs responsible for voltage regulation or protection logic.

SecureAI monitors remote-access activity, HMI command patterns and PLC responses. It would detect abnormal command timing or parameter changes that deviate from established baselines, such as bulk modifications of protection thresholds or rapid sequence execution outside maintenance windows. SecureAI then enriches alerts with affected assets, event relationships and estimated severity (Chalkias, 2024). This enables operators to take concrete mitigation actions—such as isolating the compromised access point, restoring PLC configurations or activating contingency procedures—while retaining human oversight for safety-critical decisions.

The CTI Framework structures the enriched alert, correlates it with further indicators (e.g., suspicious IPs, failed-authentication patterns, or known threat-actor TTPs) and converts it into STIX/TAXII packages supporting early detection (Art. 21), reporting workflows (Art. 23) and trusted information sharing (Art. 29) (European Parliament and Council, 2022). If deployed operationally, this structured evidence could also support internal post-incident reviews and regulator audits.

6. Discussion and Conclusions

The findings align with research showing that AI-driven anomaly detection and CTI automation enhance situational awareness and response effectiveness in OT environments (Gong & Lee, 2021; Govea et al., 2024). SecureAI reflects these approaches through integrated detection, contextual enrichment and structured intelligence dissemination. Its ability to relate events across assets and generate machine-readable intelligence reduces cognitive load for analysts and supports the structured decision-making expected in NIS2-compliant operations.

The scenario demonstrates how SecureAI supports NIS2-aligned tasks by detecting abnormal behaviour, linking events to assets and providing evidence that triggers containment steps such as isolating compromised interfaces or reverting PLC settings. The CTI Framework adds auditability by generating structured, traceable incident data that regulators or internal teams can verify. This is important because NIS2 emphasises demonstrable accountability: operators must not only act but also show why they acted and what evidence informed their decisions.

However, several limitations persist. AI-based anomaly detection and CTI automation performance depend heavily on telemetry quality, and many OT environments still lack adequate logging or network visibility—issues identified repeatedly in industry assessments (ENISA, 2024). In addition, anomaly detection may generate ambiguous alerts that require contextual understanding from operators. Organisational readiness—including training, defined escalation paths and a culture of cyber-physical safety—is therefore as critical as the technology itself. Future integration with explainable AI techniques may also help operators better interpret model outputs during high-pressure events.

This study assessed how SecureAI can support selected operational obligations of the NIS2 Directive for energy-sector OT environments. The mapping and scenario analysis show that SecureAI enhances detection, contextualisation and structured reporting, contributing to NIS2 requirements for monitoring, incident handling and information sharing. When combined with human oversight, these tools offer a practical path to operationalising NIS2 duties.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: Keenious was used for literature exploration, and ChatGPT 5.1 for drafting and refinement. Human authors verified all content.

References

Ajayi, O. O., Alozie, C. E., & Abieba, O. A. (2025). Enhancing cybersecurity in energy infrastructure: strategies for safeguarding critical systems in the digital age. *Trends in Renewable Energy*, *11*(2), 201-212.

Alevizos, L. (2025). Automated cybersecurity compliance and threat response using AI, blockchain and smart contracts. *International Journal of Information Technology*, *17*(2), 767–781. https://doi.org/10.1007/s41870-024-02324-9

Aydın, Z. (2025). Detecting Cybersecurity Threats in Digital Energy Systems Using Deep learning for Imbalanced Datasets: Deep Learning for Cyber Threat Detection in Critical Energy SCADA. *International Journal of Energy Economics and Policy*, *15*(3), 614–628. https://doi.org/10.32479/ijeep.19649

Chalkias, I. (2024). *D4.1 – Initial prototypes of the cyber-threat intelligence gathering, extraction, sharing components and AI-based solutions. [*. https://horizon-dynamo.eu/wp-content/uploads/2025/02/DYNAMO-RPT-D41-V2-0.pdf

Dragos Inc. (2025). *ICS/OT Cybersecurity Report* (Annual Year in Review No. 8; p. 56). Dragos Inc. https://pkcert.gov.pk/uploads/2025/02/Dragos-2025-OT-Cybersecurity-Report-A-Year-in-Review.pdf

ENISA. (2024, December 10). *Cyber Europe tests the EU Cyber Preparedness in the Energy Sector* [Press Release]. https://www.enisa.europa.eu/news/cyber-europe-tests-the-eu-cyber-preparedness-in-the-energy-sector

European Parliament and Council. (2022). *Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02022L2555-20221227

Gong, S., & Lee, C. (2021). Cyber Threat Intelligence Framework for Incident Response in an Energy Cloud Platform. *Electronics*, *10*(3), 239. https://doi.org/10.3390/electronics10030239

Govea, J., Gaibor-Naranjo, W., & Villegas-Ch, W. (2024). Transforming Cybersecurity into Critical Energy Infrastructure: A Study on the Effectiveness of Artificial Intelligence. *Systems*, *12*(5), Article 5. https://doi.org/10.3390/systems12050165

Ruohonen, J. (2024). *A Systematic Literature Review on the NIS2 Directive* (No. arXiv:2412.08084). arXiv. https://doi.org/10.48550/arXiv.2412.08084

Seethala, S. (2025). An AI-Driven Compliance Intelligence Platform for Continuous Monitoring and Automated Risk Assessment in Regulated CRM and ERP Systems. *International Journal of Scientific Research and Engineering Trends*, *11*(6), 2395-566X. https://doi.org/10.5281/zenodo.17938605

ECHO Early Warning System for Water Infrastructure Protection

Ilkka Tikanmäki[1,2][1] , Sara Jylhänkangas, Kitty Tapola and Jeni Awa
[1]Security and Risk Management, Laurea University of Applied Sciences, Espoo, Finland
[2]Department of Warfare, National Defence University, Helsinki, Finland

ilkka.tikanmaki@laurea.fi

Abstract: Water is a valuable natural resource and essential for human life, so protecting critical water infrastructure is vital. Through directives such as the Network and Information Security Directive (NIS2), the European Union has made it mandatory to ensure that all critical infrastructure is adequately protected. Security measures taken to protect critical infrastructure are often limited to a national scope. The ECHO Early Warning System (E-EWS) is a tool that enables communication and cooperation across national borders. It focuses on a warning system that allows people to know almost immediately what is happening, to react appropriately and to disseminate this information to trusted partners. With geopolitical tensions and cyberattacks on the rise, this work in progress employs a literature review and case analysis to identify similarities between recent cyberattacks on critical water infrastructure and to determine what E-EWS could have done in these instances. The results show that key similarities correlate with current events: political threat actors, slow detection, fear as a weapon, information silos and people as weak links. With the E-EWS, threats are detected almost immediately, allowing mitigation actions to begin much earlier. By sharing information, knowledge grows, connecting and strengthening the overall security of E-EWS users. Attackers do not care about borders, so the cybersecurity of critical water infrastructure is no longer limited to national borders. International cooperation would help reduce the burden caused by some countries' own resource constraints. Based on these results, the E-EWS plays a key role in the future protection of EU critical water infrastructure.

Keywords: Cybersecurity, Critical infrastructure, DYNAMO project, E-EWS, Information sharing

1. Introduction

Water is a critical resource essential to human life. Due to this importance, many countries categorise it as falling under the classification of critical infrastructure (European Commission, 2025; European Parliament, 2022). Critical water infrastructure can range from the supply of safe drinking water to the transportation and handling of wastewater.

In January 2023, the Network and Information Security Directive (NIS2) came into effect in the European Union (European Parliament and Council, 2022). This directive requires critical infrastructure operators to make sure that there are adequate security measures in place and that they can report any security incidents promptly to the proper authorities. As technology progresses, so does the need to incorporate more modern technology into critical infrastructure systems. Ensuring that security is a top consideration when building up critical infrastructure systems is essential (ENISA, 2017).

This paper aims to discover common characteristics from recent cyberattacks against critical water infrastructure and use those characteristics as a basis to determine the part E-EWS could potentially have to play in boosting the security of critical water infrastructure.

Section 2 examines the common recurring characteristics of cyberattacks against critical water infrastructure. Section 3 describes the ECHO Early Warning System, explaining its core functions and potential benefits before, during, and after cyberattacks. Section 4 discusses the limitations of national-only defence approaches and examines the European regulatory framework, specifically the NIS2 Directive and the Cyber Resilience Act. Section 5 concludes with a summary of findings and recommendations for future research.

2. Common Characteristics of Cyberattacks

As technology becomes increasingly prevalent in daily life, so too has the occurrence of cyberattacks. Critical infrastructure has now slowly begun to incorporate various types of technology that are providing new openings through which attackers can gain access. The aims of gaining access via cyberattacks can range from extorting money from organisations to completely shutting the infrastructure down, rendering it useless. The attack vectors also vary just as much, ranging from malware to phishing attacks (Riggs et al., 2023). When comparing recent cases of attacks on critical water infrastructure side by side, several common characteristics arose, such as political threat actors, slow detection, using fear as a weapon, information silos, and humans proving to be the weak link (Bendovschi, 2015). All these characteristics reflect what is currently happening in the world. While

[1]https://orcid.org/0000- 0001-8950-5221

there aren't many things one can do to combat the reasons why an attacker chooses to perform an attack, there are ways to mitigate the issues that relate to people. The most critical thing that can be done is to provide information and training so that cyberattacks can be known, learned from, and responded to competently (Riggs et al., 2023). These training and information gathering sessions can come in many different forms, such as workshops, webinars, and courses (ENISA, 2025).

3. ECHO Early-Warning System

The focus of the ECHO Early-Warning System (E-EWS) as a tool is to support an organisation's security operations. Security operation centres (SOC) provide a tool that allows for completely independent management of both incidents that occur and relevant cybersecurity data of their own organisation. At the same time, it allows them the capability to collaborate with others through the sharing of that same information. To ensure security, E-EWS works in conjunction with the Data Anonymisation Tool (DAT) for anonymisation and the Fine-Grained Access tool for encryption and decryption of cyber tickets.

Through E-EWS, an organisation can also have subscriptions to receive notifications on any type of information that they are interested in. An example of this would be that maybe an organisation is interested in getting a notification about the creation of a warning related to water plants. They would immediately receive a notification if this occurred within the E-EWS database (Almén et al., 2022; DYNAMO project, 2024).

E-EWS has the potential to be a benefit during all phases of an attack: before, during, and after. Before an attack occurs, E-EWS can provide organisations with knowledge of known threats and attacks. This knowledge would allow them to protect themselves better by implementing proactive measures to prevent attacks and to learn about what they could potentially be up against. Effective threat information sharing provides organisations with a deeper understanding of the threat landscape. This understanding enables the identification of affected platforms and the implementation of protective measures promptly.

During an attack, real-time sharing and collaboration via E-EWS allow organisations to manage dealing with the situation. They would easily find information relevant to mitigating any currently occurring incident. Having this kind of easy access to information and communication abilities allows organisations to feel more at ease, even when they are under attack. This helps reduce fear and human shortcomings, such as a lack of knowledge. It would also increase the speed with which incidents are noticed and dealt with. Information sharing facilitates responses against hybrid threats (Cybersecurity and Infrastructure Security Agency CISA, 2025), which seems to be what is lacking in many incidents, where detection times often range from hours to months.

After an attack, E-EWS allows sharing information about incidents. This would allow organisations to keep information for their own future use and ensure that others can learn and profit from their experiences. The system increases the degree of protection by reducing viable attack vectors across the entire network of participating organisations. When one organisation shares what happened, other organisations in the network can adapt and become more secure against similar attacks.

4. National-Only Defence

The EU has created policies against cyber threats, such as the NIS2 Directive and the Cyber Resilience Act (European Commission. Joint Research Centre, 2024; European Parliament and Council, 2022). Policies alone cannot guarantee compliance or implementation. In recent years, there has been a rise in cybersecurity attacks across Europe, and when attackers decide to attack, they ignore borders. Being unable to communicate safely and effectively with a neighbouring country that is under attack could have devastating consequences. A unified EU has more capabilities to provide the right amount of help where it matters. Sharing information across the union will provide better defence against cyberattacks than a policy could accomplish by itself.

5. Conclusions

The increase in cyberattacks will remain. Even though attackers are becoming more creative and adapting with the times, there are common characteristics that many cyberattacks often share. Some of these things, such as the motive behind an attack, are things that cannot be controlled or mitigated. Other things, such as slow detection, fear, information silos, and human weakness, are things that can be controlled. The implementation of an early warning system, such as E-EWS, is a benefit to the cybersecurity of critical water infrastructure. As attackers themselves are not held back by borders, neither should the cybersecurity of critical water infrastructure be held back by them. All the characteristics that can be mitigated relate to people and something that they are lacking, whether it be information, education, or time. E-EWS provides ways to help solve those

issues. When we investigate Finland's current situation, we can see that it is a good representation of the inherent flaws of a national-only-based defence system, but we are also able to see the positives, namely, that Finland understands the need for international cooperation and communication going into the future. Further research is necessary to determine the practicality of implementing E-EWS as part of the cybersecurity of critical water infrastructure. It would be necessary to research the best approach for each country and organisation's needs. Small-scale simulations might also look into the viability of E-EWS on a national and international scale.

Acknowledgements

This study has received funding from the European Union project DYNAMO, the European Union's Horizon 2020 research and innovation programme under the grant agreement no. 101069601. The views expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: Keenious was used for literature exploration, and ChatGPT 5.1 for drafting and refinement. Human authors verified all content.

References

Almén, C., Hagström, N., & Rajamäki, J. (2022). ECHO Early Warning System as a Preventive Tool against Cybercrime in the Energy Sector. *Information & Security: An International Journal*, *53*(1), 11–20. https://doi.org/10.11610/isij.5301

Beek, K. (2025). *Russian Hacktivists Take Aim at Polish Power Plant, Again*. Dark Reading. https://www.darkreading.com/cyberattacks-data-breaches/russian-hacktivists-polish-power-plant-attack

Bendovschi, A. (2015). Cyber-Attacks – Trends, Patterns and Security Countermeasures. *Procedia Economics and Finance*, *28*, 24–31. https://doi.org/10.1016/S2212-5671(15)01077-1

Cybersecurity and Infrastructure Security Agency CISA. (2025). *Information Sharing*. Cyber Threats and Advisories. https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing

Doyle, A. (2025a, February 27). *£4.5 Million Ransomware Bill for Southern Water After Black Basta Attack* [Security Spotlight]. https://dailysecurityreview.com/security-spotlight/black-basta-ransomware-costs-southern-water-4-5-million/

Doyle, A. (2025b, August 22). *Norway Attributes Dam Cyberattack to Russian Hackers—Cybersecurity* [Cybersecurity]. https://dailysecurityreview.com/cyber-security/norway-attributes-dam-cyberattack-to-russian-hackers/

DYNAMO project. (2024). *D4.1 – Initial prototypes of the cyber-threat intelligence gathering, extraction, sharing components and AI-based solutions. [*. https://horizon-dynamo.eu/wp-content/uploads/2025/02/DYNAMO-RPT-D41-V2-0.pdf

ENISA. (2017). *Communication network dependencies for ICS/SCADA Systems*. European Union Agency for Network and Information Security (ENISA). https://www.enisa.europa.eu/sites/default/files/publications/WP2016%203-1%202%20ICS%20SCADA%20Dependencies.pdf

ENISA. (2025). *NIS2 Technical Implementation Guidance*. European Union Agency for Cybersecurity (ENISA). https://doi.org/10.2824/2702548

European Commission. (2025). *Critical infrastructure resilience at EU-level—Migration and Home Affairs* [Policy]. Critical Infrastructure Resilience at EU-Level. https://home-affairs.ec.europa.eu/policies/internal-security/counter-terrorism-and-radicalisation/protection/critical-infrastructure-resilience-eu-level_en

European Commission. Joint Research Centre. (2024). *Cyber resilience act requirements standards mapping: Joint Research Centre & ENISA joint analysis.* (Analysis No. EUR 31892; p. 69). Publications Office of the European Union. https://data.europa.eu/doi/10.2760/905934

European Parliament. (2022). *The resilience of critical entities*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022L2557

European Parliament and Council. (2022). *Directive (EU) 2022/2555 on measures for a high common level of cybersecurity across the Union (NIS2 Directive)*. Official Journal of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02022L2555-20221227

Holm Security. (n.d.). *NIS & NIS2 in Norway* [Resources]. NIS2 & the EEA. Retrieved January 9, 2026, from https://www.holmsecurity.com/nis2-in-norway

Khomenko, I. (2025, August 19). *Russian Hackers Breach Polish Hydropower Plant in Major Cyberattack*. UNITED24 Media. https://united24media.com/latest-news/russian-hackers-breach-polish-hydropower-plant-in-major-cyberattack-10882

Pukėnaitė, I. (2025, August 19). *Russian hackers take control of Polish hydropower plant, turbines disrupted*. Cybernews. https://cybernews.com/cybercrime/russian-hackers-target-polish-hydropower-plant-again/

Riggs, H., Tufail, S., Parvez, I., Tariq, M., Khan, M. A., Amir, A., Vuda, K. V., & Sarwat, A. I. (2023). Impact, Vulnerabilities, and Mitigation Strategies for Cyber-Secure Critical Infrastructure. *Sensors*, *23*(8), 4060. https://doi.org/10.3390/s23084060

Automated Exploit Chain Modeling and Analysis

Thomas Wahl[1], Nicholas White[1], Guang Jin[1], Sukarno Mertoguno[2], Kevin Stevens[2] and Froy Maldonado[2]

[1]Trusted Science and Technology, Inc., Rockville, MD, USA

[2]Georgia Institute of Technology, Atlanta,GA, USA

twahl@trustedst.com (corresponding author)

Abstract: We describe early-stage research and tool development efforts to formally model and analyze *exploit chains*. These are sequences of exploits carefully crafted by an attacker to achieve an elaborate end-goal, such as an escalation of privileges of the executing thread. In this work, we are taking a systematic approach to constructing *formal models* of exploit chains in the form of finite-state machines, which are then converted into constraint-based semantic representations or timed automata, in order to analyze chains against metrics such as effectiveness, ease of reproduction, and stability under system variations.

Keywords: Exploits, Exploit chains, Vulnerabilities, Formal modeling, Analysis

1. Introduction and Background

While the concept of security exploits has been the subject of extensive studies, exploit *chains* have enjoyed less attention in the research community; existing work tends to be ad-hoc. What makes such chains challenging to recognize and understand is their trademark behavior of crossing diverse layers of a complex computational system (such as a mobile phone), including the device firmware, driver code, operating system code (e.g., Android kernel routines), user-level code, and application code (e.g., web browsers).

1.1 Exploit Chains

These are sequences of exploits designed so that one exploit achieves a sub-goal of the chain (such as escaping a security sandbox or the execution of attacker-provided code) that serves as the gateway to the next exploit. The final exploit accomplishes the ultimate attacker goal, such as an escalation of privileges of the executing thread, or the exfiltration of some data meant to be inaccessible to standard users. Individual exploits are typically enabled by *vulnerabilities* in the underlying code, such as dereferencing a pointer to freed memory.

Consider the example of an exploit chain capable of remotely creating a root shell on a wide range of Android devices (Gong, 2020). The chain connects three exploits, based on three vulnerabilities, which target diverse software components and span multiple layers of an Android device:

- A memory overwrite access vulnerability (CVE-2019-5877) in a JavaScript engine;
- A use-after-free vulnerability (CVE-2019-5870) used to enable executing a shell system command; and
- A vulnerability in the Qualcomm GPU Linux kernel driver (CVE-2019-10567) used to switch into boot mode that results in a root shell, and reboot.

1.2 Formal Methods

This is an umbrella term for a collection of techniques that represent computational systems as *formal models*, such as labeled graph structures or algebraic constraints, and then use advanced graph search techniques or mathematical logic, respectively, to establish properties of these models. Their key strength is that, due to their algorithmic and logical rigor, they can provide guarantees of the correctness of the established model properties. A downside is that the model representations (i.e., graphs, constraint sets, logical formulas) are often far away from mainstream programming languages (requiring extensive abstraction), and that the analysis techniques are expensive and may require non-trivial human expertise to use effectively.

2. Modeling and Analysis of Exploit Chains: An Approach

Figure 1 shows an overview of our approach. Our project is not about detecting or root-causing exploit chains, but about modeling and analyzing known chains, for the (future) purposes of diversification and extension. We therefore start with an exploit chain given as executable code, as well as sources for the attack code, if available. We first apply *static and dynamic analyses* to extract a finite-state machine (FSM) model from the given description; we refer to this stage as *exploit chain ingestion*. The FSM as a model is generic in nature and lends itself to translation into different back-end analysis technologies: The figure shows the steps to extract either a

symbolic model, to be processed with a constraint solver, or a state-based model using timed automata, processed by the UPPAAL model checker (Bengtsson et al, 1996).

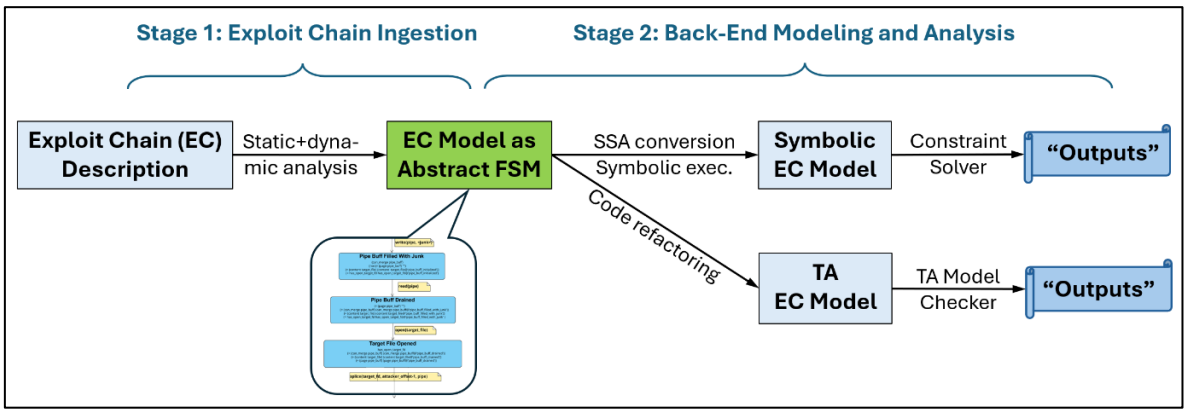


Figure 1: Overview of the two-stage exploit chain (EC) modelling and analysis process. An abstract FSM is constructed as intermediate representation. Obtaining a symbolic EC model requires converting the code into single static assignment (SSA) form, followed by symbolic execution

2.1 Exploit-Chain Modeling as a Finite-State Machine

The extraction of the FSM model employs both dynamic and static analysis techniques, at different stages of the ingestion algorithm, which we sketch in the following.

- **Determine a set of events that mark major milestones in the execution of the chain.** This step requires some high-level insights from a human expert into the exploit chain and is therefore done largely manually. Technically, the events are function calls in the various source code sections traversed by the chain. In addition, the events set includes memory allocation and deallocation statements, and thread spawns or joins.
- **Place log message instructions into the code at the event locations, and rebuild the chain code.** Challenges to automating this step include managing the various source code forms that exploit chains traverse, such as C/C++, Java, JavaScript, and shell scripts, all with different instrumentation requirements.
- **Execute the instrumented chain code, multiple times if necessary, until the chain has met its goal.** The execution generates a trace of the chain events previously identified as milestones. Tracing an execution of a full exploit chain targeting Android phones is very challenging: Such chains are bound to specific Android OS versions and other phone-specific software. We thus need either a mechanism to trace code running on a physical phone, or an Android OS emulator; we use both forms of tracing. The events serve as *states* of the FSM to be constructed.
- **Assign an execution semantics to the edges of the FSM.** This is accomplished by symbolically executing the code between two marker locations, defining the transition semantics via formulas that describe the change of state. While symbolic execution can be expensive, note that the trace dynamically constructed *focuses* the symbolic execution to a specific, thin slice of the overall code. The slice consists mostly of the path followed by the chain, but also preconditions checked on the way whose violation leads to termination of the attack. These checks are critical for assessing the robustness of the chain.

2.2 Exploit-Chain Analysis via Constraint Solving and Model Checking

The back-end analysis involves translating the FSM model into the input language of one of the available analysis tools, and run that tool. The table below compares the two tool pipelines that we support.

EC modeling language:	**Symbolic expressions**	**State transition system**
Analysis tools:	Constraint solvers: Z3, CVC4, …	(Explicit-state) Model checkers: UPPAAL, SPIN
Advantages compared to the other tool set:	Can analyze chain behavior under side conditions Can reverse-engineer inputs for a given effect	Good at handling multi-threaded scenarios Timed automata can track timing dependencies

An overview of the two pipelines is shown in Figure 2. The upper pipeline translates the FSM into the SMT ("Satisfiability modulo theories") language widely used by constraint solvers. The SMT model is then passed to a solver; we use Z3 (de Moura et al., 2008). The lower pipeline translates the FSM into the Timed Automaton (TA) input language of the UPPAAL model checker (Bengtsson et al., 1996). In both cases, the translation also takes attacker inputs and other configuration parameters as input, as well as a property specification. The property might express that the final state of the running exploit chain has given one of the participating threads root privileges, or that some content of a read-only database has been copied into a user (= attacker) variable.

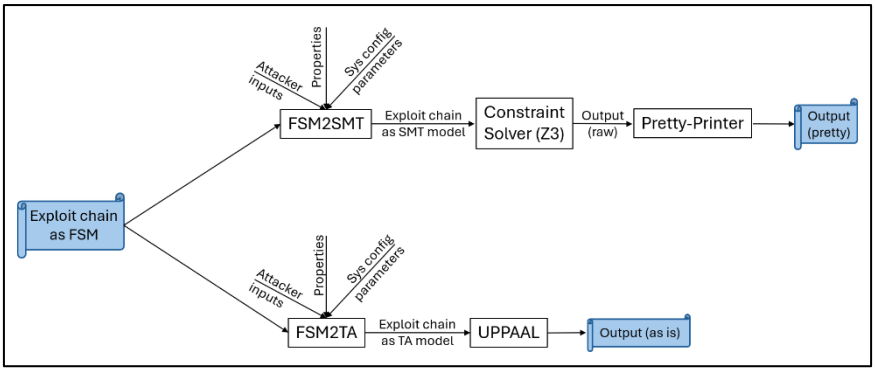


Figure 2: Two back-end analysis pipelines operating on top of the FSM

The decision which analysis pipeline to use is largely manual at this time but is of course heavily informed by the capabilities of the analysis tool. For example, an attack involving multiple threads will likely be processed using the UPPAAL-based pipeline.

2.3 Preliminary Results

We have so far created an initial implementation of the ingestion of a runnable exploit chain into an analysis model, and we have a decent implementation of the subsequent assessment of chain properties. We have used the implementation to analyze exploit chains at the kernel level, such as the infamous Dirty Pipe file-overwrite exploit (https://dirtypipe.cm4all.com), and Android exploits targeting mobile phones. We present preliminary results using the Dirty Pipe.

Property	Result	Interpretation
P1: E <> M0.end	Satisfied	We can reach the end state
P2: A [] M0.end => imply M0.main_1_ret_val == 0	Satisfied	Whenever the end state is reached, the Dirty Pipe has succeeded.
P3: Pr[<= 100] (<> M0.end)	P=[0.950056,1]	

The table above shows three Dirty Pipe properties P1..P3 in temporal-logic notation (as accepted by the UPPAAL model checker), the checker outcome, and an explanation. For example, P1 expresses that, along some path (E), machine M0 will eventually (<>) reach the end state. P3 asks: for paths up to 100 time units long, what is the probability of reaching the end state? In some cases it is necessary to introduce *ghost variables* into the model, solely used to express these properties. For example, the Dirty Pipe aims to overwrite a file that the attacker does not have write access to. Rather than modeling the complete Android filesystem, we model files as strings, embellished with a file access permission attribute (and others). The overwrite outcome of the Dirty Pipe is tested in the model by string comparison; if a change is observed, variable ret_val of function main_1 is set to 0 (property P2).

Property	States explored	CPU time (ms)	Virtual memory (kB)	Resident memory (kB)
P1	11441	56	154244	107607
P2	11442	61	154925	108135
P3	1647504	242	155141	108780

The table above quantifies the cost of analyzing the properties over the timed-automaton model, using UPPAAL. We see that, while the running times are very small, the memory requirements are sizable. However, this is in part due to a large start-up memory cost for state machine representation. We have run separate experiments,

using artificial examples, demonstrating that the memory requirements grow reasonably against growing model size.

3. Future Work

In addition to significant development work on the current preliminary implementation (such as the collaboration of the dynamic and static analyses for ingesting the exploit chains into an FSM), future plans include developing techniques for automating the repair of chains (if necessary), in the face of

- **incidental system changes:** generic kernel or software updates, and
- **system hardening:** patching of a vulnerability involved in the chain.

More ambitious yet is the goal of *forecasting* exploit chains: given a system and an anticipated software upgrade, what new opportunities for exploit chain creation does the upgraded system offer? Answers to such questions may furnish insights into both adversarial use of exploit chains, as well as defense against them. Both chain repair and forecasting likely offer several solutions per scenario, rather than a "unique best". We may consider Multi-Criteria Decision Making approaches under uncertainty to present the user with an informed choice.

Acknowledgements

This research was developed with funding from the US Defense Advanced Research Projects Agency (DARPA) under the "INGOTS: Intelligent Generation of Tools for Security" program.

Disclaimer: The views, opinions, and/or findings expressed are those of the author(s) and should not be interpreted as representing the official views or policies of DARPA or the U.S. Government. Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).

AI Declaration: AI tools were not used in the creation of this paper.

Ethics Declaration: Ethical clearance was not required for this research.

References

de Moura, Leonardo Mendonca and Bjorner, Nikolaj (2008) "Z3: An Efficient SMT Solver", *Conference on Tools and Algorithms for the Construction and Analysis of Systems* (TACAS). DOI: https://doi.org/10.1007/978-3-540-78800-3_24

Gong, Guang (2020) "TiYunZong: An Exploit Chain to Remotely Root Modern Android Devices", Black Hat USA.

Bengtsson, Johan and Larsen, Kim and Larsson, Fredrik and Pettersson, Paul and Yi, Wang (1996) "UPPAAL—a tool suite for automatic verification of real-time systems", DIMACS/SYCON workshop on Hybrid systems III: Verification and Control. DOI: https://doi.org/10.1007/BFb0020949

Late Submissions

Cyberbiosecurity in Action: Securing the Biological Battlespace from Digital Subversion

Adya Daruka[1], Anusha Nigam[2] and Shreyas Kumar[3]
[1]University of Illinois Urbana Champaign, USA
[2]Centennial High School, Frisco, USA
[3]Texas A&M University, College Station, USA

adaruka2@illinois.edu
anusha.nigam.0101@gmail.com
shreyas.kumar@tamu.edu

Abstract: Cyberbiosecurity is an emerging discipline at the intersection of cybersecurity and life sciences, as digital dependence across genomics, laboratory automation, biomanufacturing, and digital health expands the attack surface of the bioeconomy. This paper synthesizes peer-reviewed literature, government reports, and public incident investigations to map cyber-biological threats, document observed incidents (e.g., healthcare ransomware and documented proofs-of-concept for DNA-encoded exploits), and identify systemic gaps in preparedness, workforce capacity, and governance. The literature indicates recurring vulnerabilities in healthcare and biotechnology supply chains, demonstrated operational impact from ransomware incidents, and credible research demonstrating new attack vectors such as malware encoded in synthetic DNA sequences. Drawing on these sources, we present the Integrated Cyberbiosecurity Defense Framework (ICDF) - a proposed multi-layered strategy combining bio-risk assessment, AI-driven detection, digital-twin simulation, fusion centers for cross-disciplinary response, and workforce development - and recommend policy, technical, and organizational steps for immediate adoption.

Keywords: Cyberbiosecurity, Biological battlespace, Bio-digital threats, Genomic data security, Synthetic biology cybersecurity, Digital health risk

1. Introduction

The convergence of biological research and digital technology has created new opportunities and vulnerabilities. As life sciences increasingly depend on computational methods, digital databases, and automated systems, they are exposed to cyber threats beyond the scope of traditional biosecurity. This fusion introduces risks that require urgent interdisciplinary attention, giving rise to the emerging field of cyberbiosecurity (Murch et al., 2018; Peccoud, 2018). Cyberbiosecurity focuses on protecting biological data, bioinformatics platforms, manufacturing systems, and healthcare technologies from threats such as hacking, ransomware, and espionage. Compromises in this domain can endanger intellectual property, public health, and national security. Yet, while cybersecurity and biosecurity have developed independently, their intersection remains underexplored. Cybersecurity frameworks rarely address the unique properties of biological data or the complexities of biotech infrastructure, while biosecurity practices focus largely on physical containment without integrating digital risks. This study addresses these gaps by examining cyberbiosecurity challenges across biotechnology, healthcare, and biomanufacturing. It identifies vulnerable assets, such as genomic databases and laboratory automation systems (Nix et al., 2021), and highlights low awareness among stakeholders and insufficient organizational preparedness. To respond, we propose a comprehensive framework combining policy development, cross-disciplinary collaboration, and workforce training. By illuminating vulnerabilities at the nexus of biology and cyberspace, this work emphasizes the urgent need for proactive strategies to secure the bioeconomy and build resilience in an increasingly digital life sciences sector.

2. Background

The digitization of biological research has fundamentally transformed life sciences, enabling faster discoveries, scalable bioengineering, and personalized healthcare. Genome sequencing, synthetic biology, and bioinformatics pipelines now rely almost entirely on digital platforms for storing, analyzing, and transmitting sensitive biological data. Unfortunately, this dependence introduces vulnerabilities that traditional biosecurity frameworks never anticipated. Genomic data presents unique challenges. Unlike typical personally identifiable information, genetic data is immutable and reveals extensive insights about individuals and entire populations. Breaches can compromise privacy, affect medical treatment and insurance decisions, and carry consequences that last for generations (Liu, 2019; Weiss, 2019). The shift toward cloud-based genomic platforms further amplifies risks of unauthorized access, data manipulation, and loss (Zhang and Wang, 2020).

Laboratory automation and biomanufacturing add another layer of cyber-physical risk. Robotics, IoT devices, and computer-controlled bioreactors improve efficiency but dramatically expand the attack surface for adversaries - from cybercriminals to nation-states (Kim, 2021; Li et al., 2021). Real-world incidents underscore these dangers: the 2017 WannaCry ransomware attack disrupted patient care across the UK's NHS (National Audit Office, 2018), while 2019 research demonstrated that malware could actually be encoded into DNA sequences (Ney et al., 2017; Brumfiel, 2020), blurring the line between biological and digital threats in disturbing ways. Policy responses remain fragmented. Agencies like HHS and DHS have issued sector-specific guidelines (U.S. Department of Health and Human Services, 2018; U.S. Department of Homeland Security, 2019), but no unified framework governs cyber risks in biology. Without standardized baselines, organizations rely on inconsistent, ad hoc measures.

Cultural barriers compound these technical challenges. Life scientists often lack cybersecurity training, while security professionals typically have limited understanding of biological workflows. Bridging this divide is essential. Cyberbiosecurity is not merely a theoretical concern - it represents an urgent operational necessity. This study synthesizes documented evidence of awareness gaps, preparedness challenges, and evolving threats across biotechnology, healthcare, and cybersecurity sectors while proposing concrete interventions to support resilient, scalable practices.

3. Related Work

Cyberbiosecurity, though a young discipline, has gained traction over the past decade. Early contributions by Murch et al. (2018) and Peccoud (2018) introduced the term and emphasized collaboration between cybersecurity experts and life scientists to protect biological assets. Subsequent work has mapped specific threats: Hu et al. (2020) showed how synthetic DNA ordering systems could embed malicious code, while Nix et al. (2021) highlighted risks in laboratory automation where compromised robots or sequencing tools could corrupt or steal data. Together, these studies underscore the seriousness of cyber risks in life sciences.

Healthcare cybersecurity research offers parallel insights. Hospitals have been frequent ransomware targets, with consequences for patient safety (Kruse et al. 2017). The rise of electronic health records has improved care but also created lucrative targets for data theft (Smith, 2018; Goodman, 2020). These findings demonstrate how disruptions in bio-related systems can carry life-threatening or economically significant consequences.Despite these advances, empirical studies on cyberbiosecurity awareness and preparedness remain limited. Most prior work is conceptual, focusing on hypothetical risks or technical fixes rather than stakeholder practices in biotechnology and healthcare. Current cybersecurity standards, including ISO/IEC 27001 and the NIST Cybersecurity Framework (National Institute of Standards and Technology, 2020), offer general guidance but lack bio-specific controls, while biosecurity frameworks like the Federal Select Agent Program address physical hazards while overlooking digital threats. This regulatory gap reinforces the need for dedicated cyberbiosecurity policies.

Recent interdisciplinary workshops and whitepapers from organizations such as AAAS and DHS (U.S. Department of Homeland Security, 2019; Office of Science and Technology Policy, 2022) have called for cross-sector collaboration, but concrete models for integrating cybersecurity into biological research remain underdeveloped. This study extends the literature by synthesizing documented vulnerabilities and proposing a comprehensive framework that integrates cybersecurity, biosecurity, and organizational perspectives to strengthen resilience in the bioeconomy.

4. Methodology

This paper employs a systematic literature review and threat analysis methodology to synthesize the current state of cyberbiosecurity knowledge and propose a comprehensive defense framework. Our approach consisted of three phases:

Literature Collection. We conducted searches across peer-reviewed databases (PubMed, IEEE Xplore, Web of Science) and grey literature sources (government reports, white papers, technical documentation) using search terms including "cyberbiosecurity," "DNA malware," "laboratory automation cybersecurity," "healthcare ransomware," "bioeconomy safeguarding," and related phrases. We prioritized: (a) peer-reviewed research documenting cyber-biological attack vectors or defenses, (b) authoritative government reports on biosafety and cybersecurity, and (c) audited incident investigations from healthcare and biotechnology sectors.

Thematic Analysis. We extracted recurring themes from the literature including awareness gaps, supply-chain vulnerabilities, novel attack vectors, workforce deficits, and regulatory gaps. We analyzed documented incidents

(e.g., WannaCry NHS attack) and proof-of-concept research (e.g., DNA-encoded malware) to understand feasibility and impact of various threat scenarios.

Framework Development. Based on identified gaps and documented vulnerabilities, we synthesized findings into the Integrated Cyberbiosecurity Defense Framework (ICDF). This framework represents a conceptual architecture integrating best practices from cybersecurity, biosecurity, and risk management literature, adapted specifically for cyber-biological convergence challenges.

Limitations. This study is limited to publicly available literature and may be subject to publication bias. The rapidly evolving nature of cyberbiosecurity means new threats and defenses emerge continuously. The proposed framework has not been empirically validated and represents a theoretical contribution requiring future testing and refinement.

5. Results

Our review reveals several consistent themes across the cyberbiosecurity literature:

Awareness and Preparedness Gaps. Multiple policy reports and reviews document insufficient integration of bio-specific security measures into organizational information security programs. The National Academies' Safeguarding the Bioeconomy report highlights governance gaps and emphasizes the need for workforce development across life sciences and cybersecurity communities (NASEM, 2020). Peer-reviewed commentaries consistently identify a shortage of interdisciplinary training programs and professionals with both biological and cybersecurity expertise (Richardson et al., 2019).

Healthcare Incident Impact. Documented cyber incidents in healthcare demonstrate concrete operational harms. The 2017 WannaCry ransomware incident affecting the UK's NHS resulted in cancelled procedures, disrupted diagnostics, and prolonged service impacts (National Audit Office, 2018; Martin et al., 2017; Landi, 2019). European cybersecurity analyses document similar patterns across healthcare systems, illustrating how cyber incidents cascade into clinical harm (ENISA, 2021).

Supply-Chain Vulnerabilities. Cyberbiosecurity literature consistently identifies biological supply chains - including DNA synthesis providers, reagent vendors, outsourced laboratory services, and cloud bioinformatics platforms - as major systemic vulnerabilities where compromises can propagate through research and production pipelines (Richardson et al., 2019; ENISA, 2021).

Novel Attack Vectors. Experimental research demonstrates new classes of threats. Ney et al. (2017) demonstrated the conceptual feasibility of encoding exploit payloads into DNA sequences that, when processed by vulnerable bioinformatics pipelines, could trigger execution of malicious code. Additional research documents risks associated with compromised laboratory automation systems including robots and sequencers (Nix et al., 2021; Peccoud, 2018; Richardson et al., 2019).

Nation-State Targeting. Defense-oriented analyses indicate that intelligence communities view biological research infrastructure as strategic targets for espionage and capability theft (Murch et al., 2018; ENISA, 2021). Documented intrusion campaigns affecting research institutions and vaccine development programs illustrate this threat category.

Regulatory Gaps. Existing frameworks including the NIST Cybersecurity Framework (National Institute of Standards and Technology, 2020) and HIPAA (U.S. Department of Health and Human Services, 2018) provide general control baselines but lack biology-specific controls for genomic database integrity, sequencing pipeline security, and synthetic biology platforms (NASEM, 2020; HHS, 2018).

Economic Impact. Incident investigations and sector assessments show that cyber incidents in health and biotechnology carry significant costs due to regulatory penalties, clinical disruption, and intellectual property loss (Martin et al., 2017; ENISA, 2021).

6. Findings

Based on our literature synthesis, we identify the following critical challenges that current cybersecurity approaches fail to adequately address:

6.1 Novel Threat Landscape

The literature documents a transformed threat environment where biological materials themselves can become cyber attack vectors. Research demonstrating DNA-encoded malware (Ney et al., 2017) represents a paradigm shift requiring new defensive approaches. Additionally, emerging AI capabilities in synthetic biology create new possibilities for both threat development and defense (Hu et al., 2020; Zhang and Wang, 2020).

6.2 Inadequate Bio-specific Controls

Organizations consistently lack security controls tailored to biological systems. Standard cybersecurity frameworks do not address unique challenges such as genomic data protection, synthetic biology platform security, or biomanufacturing resilience (Peccoud, 2018; Peccoud and Peccoud, 2020; NASEM, 2020).

6.3 Supply Chain as Attack Surface

The interconnected nature of biological research and manufacturing creates cascading vulnerability chains. Literature documents risks across DNA synthesis platforms, biological reagents, and research collaboration networks (Richardson et al., 2019; Berger, 2021).

6.4 Strategic Targeting

Government and defense analyses indicate that sophisticated actors increasingly target cyberbiosecurity assets, including genomic databases, vaccine research programs, and biodefense laboratories (Murch et al., 2018).

6.5 Regulatory Insufficiency

Current regulatory frameworks including HIPAA (U.S. Department of Health and Human Services, 2018), the NIST Cybersecurity Framework (National Institute of Standards and Technology, 2020), and international biosecurity conventions do not adequately address cyber-biological convergence risks, creating uncertainty for organizations attempting compliance (Kruse et al., 2017; NASEM, 2020).

6.6 Workforce Deficits

Multiple sources identify cyberbiosecurity expertise as a critical capability gap. The interdisciplinary nature of required knowledge - spanning cybersecurity, biology, and policy - creates recruitment and training challenges. Literature documents systematic failures to develop cyberbiosecurity curricula in academic institutions (Richardson et al., 2019).

6.7 Technology Integration Challenges

Reviews note fundamental incompatibilities between traditional cybersecurity technologies and biological research environments. Legacy laboratory systems, specialized biological databases, and real-time bioprocess monitoring create unique security challenges that conventional tools cannot adequately address (Hu et al., 2020; Hardin, 2021; Miller, 2020).

6.8 Resource Requirements

Incident analyses suggest cyberbiosecurity breaches carry higher costs than traditional cybersecurity incidents due to regulatory penalties, intellectual property loss, and operational disruption (Martin et al., 2017; Landi, 2019; Smith, 2018). This necessitates greater investment in specialized expertise, technology, and processes.

7. Discussion

7.1 Framework Rationale

Our literature synthesis reveals that traditional cybersecurity approaches are insufficient for addressing cyber-biological convergence challenges. Current frameworks lack specialized controls for biological data protection, synthetic biology security, and biomanufacturing resilience (Murch et al., 2018; Richardson et al., 2019). The documented emergence of DNA-based attack vectors (Ney et al., 2017) and AI-augmented biological threats (Hu et al., 2020) represents a paradigm shift requiring purpose-built defensive strategies.We propose the Integrated Cyberbiosecurity Defense Framework (ICDF) as a multi-layered approach that addresses these unique challenges through six interconnected components designed to provide comprehensive protection across the biological domain.

7.2 Framework Components

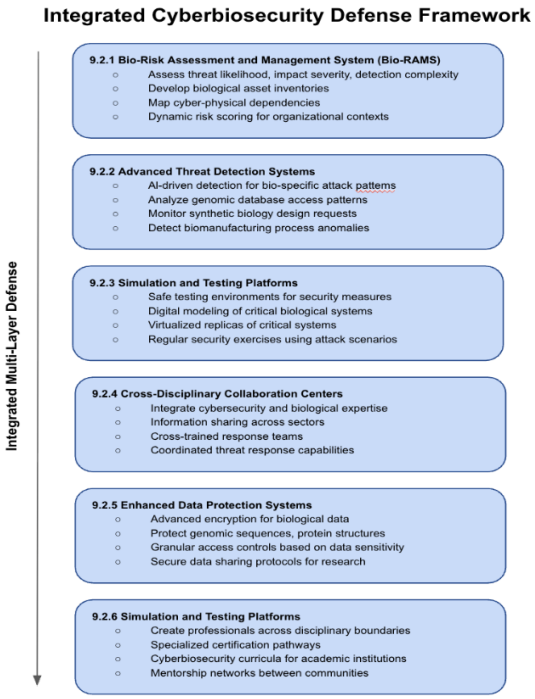


Figure 1: The Integrated Cyberbiosecurity Defense Framework (ICDF) showing six interconnected components for protecting biological systems from cyber threats

7.2.1 Bio-risk assessment and management system (Bio-RAMS)

Building on traditional risk assessment methodologies (NIST, 2020), Bio-RAMS extends these approaches to address cyber-biological threats specifically. This component assesses threat likelihood, impact severity, and detection complexity across genomic databases, synthetic biology platforms, and biomanufacturing systems. The system provides dynamic risk scoring adapted to organizational contexts and emerging threats.

Organizations should develop comprehensive biological asset inventories, map cyber-physical dependencies, and establish risk matrices accounting for both digital and biological consequences of security failures.

7.2.2 Advanced threat detection systems

Drawing on documented AI capabilities in threat detection (Hu et al., 2020), this component proposes specialized detection systems designed to identify bio-specific attack patterns. These systems analyze genomic database access patterns, synthetic biology design requests (Berger, 2021), and biomanufacturing process anomalies to identify potential attacks.

Deployment requires behavioral analytics tailored to biological workflows, baseline profiles for normal biological research activities, and alert mechanisms for anomalous patterns indicating potential compromise.

7.2.3 Simulation and testing platforms

This component develops safe testing environments for security measures without risking actual biological assets. Digital modeling of critical biological systems (Miller, 2020) enables organizations to evaluate security controls and incident response procedures in controlled settings.

In practice, organizations create virtualized replicas of critical systems, conduct regular security exercises using documented attack scenarios, and refine defenses based on simulation outcomes.

7.2.4 Cross-disciplinary collaboration centers

Addressing documented collaboration deficits (Richardson et al., 2019), this component establishes specialized centers integrating cybersecurity and biological expertise. These centers facilitate information sharing between biotechnology, healthcare, cybersecurity, and defense sectors while providing coordinated threat response capabilities.

Implementation requires establishing information sharing agreements, creating cross-trained response teams, developing common threat terminology, and conducting joint exercises across sectors.

7.2.5 Enhanced data protection systems

Recognizing the unique characteristics of biological data and long-term protection requirements, this component proposes advanced encryption and access control systems specifically designed for biological information including genomic sequences, protein structures, and bioinformatics workflows (Weiss, 2019; Zhang and Wang, 2020).

Organizations must implement strong encryption for biological data at rest and in transit, establish granular access controls based on data sensitivity, and develop secure data sharing protocols for research collaboration.

7.2.6 Workforce development program

This component addresses documented expertise gaps (Richardson et al., 2019) through comprehensive training initiatives creating professionals capable of operating across traditional disciplinary boundaries. The program includes specialized certification pathways, cross-training opportunities, and research collaboration mechanisms.

A practical approach involves developing cyberbiosecurity curricula for academic institutions, creating professional certification programs, establishing mentorship networks between cybersecurity and biological sciences communities (Peccoud and Peccoud, 2020), and supporting interdisciplinary research initiatives.

7.3 Implementation Roadmap

We propose a phased implementation approach spanning four years. Phase 1, covering the first twelve months, focuses on foundation building through organizational cyberbiosecurity assessments that identify current vulnerabilities and gaps. During this phase, organizations would develop comprehensive biological asset inventories cataloging genomic databases, laboratory automation systems, and biomanufacturing infrastructure. Baseline security controls would be established to address immediate risks, while workforce training programs would begin building internal expertise. The expected outcome of Phase 1 is basic cyberbiosecurity awareness across the organization and initial implementation of fundamental security controls.

Phase 2, spanning months 13 through 24, emphasizes capability development by deploying advanced detection systems tailored to biological workflows and attack patterns. Organizations would establish simulation environments for safe testing of security measures and incident response procedures without risking actual biological assets. This phase includes creating cross-sector collaboration mechanisms to enable information sharing with peer organizations in healthcare, biotechnology, and cybersecurity sectors. Enhanced data protection systems specifically designed for biological information (Weiss, 2019; Liu, 2019) would be implemented during this period. The expected outcome of Phase 2 is functional cyberbiosecurity capabilities with measurably improved threat detection and response times.

Phase 3, covering months 25 through 36, focuses on integration and optimization by bringing together all framework components into a cohesive defense system. Organizations would conduct comprehensive security exercises including red team assessments and simulated attack scenarios to test their integrated capabilities. Processes would be refined based on operational experience and lessons learned from exercises and real incidents. Workforce development programs would expand to build deeper expertise and create internal champions for cyberbiosecurity. The expected outcome of Phase 3 is a mature cyberbiosecurity program with coordinated defense capabilities across all organizational functions.

Phase 4, spanning months 37 through 48, develops advanced capabilities including deployment of automated response systems where appropriate to reduce response times for known threat patterns. Organizations would establish predictive threat intelligence capabilities (Choucri et al., 2019) to anticipate emerging risks before they manifest. Full cross-sector information sharing would be achieved, enabling real-time threat awareness across the broader community. Continuous improvement and adaptation mechanisms would be embedded to ensure the program evolves with the threat landscape. The expected outcome of Phase 4 is an advanced, adaptive cyberbiosecurity posture capable of addressing both current and emerging threats.

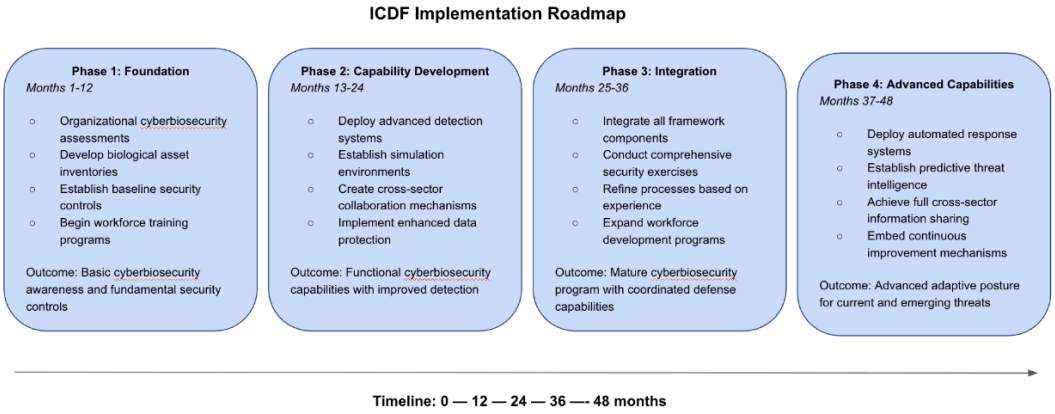


Figure 2: Four-phase implementation roadmap for deploying the ICDF over 48 months, from foundation building to advanced capabilities

7.4 Validation Requirements

The proposed framework requires empirical validation through multiple assessment approaches. Pilot implementations in diverse organizational contexts - academic research institutions, commercial biotechnology firms, healthcare facilities, and government laboratories - will test the framework's applicability across different operational environments. Red team exercises using documented attack vectors from the literature (Ney et al., 2017; Hardin, 2021) can evaluate the framework's effectiveness against known threats.

Longitudinal assessment of security improvements should track changes in vulnerability levels, incident frequency, and response effectiveness over extended periods. Cost-benefit analysis of implementation can quantify resource requirements against reduced incident risk and operational benefits. Finally, stakeholder feedback from biotechnology, healthcare, and cybersecurity sectors will ensure the framework addresses real-world needs and constraints faced by practitioners.

7.5 Expected Outcomes and Limitations

If successfully implemented, the ICDF framework should improve detection of cyber-biological threats through specialized monitoring systems attuned to biological workflow anomalies. The framework will enhance organizational preparedness through workforce development programs creating internal expertise spanning both cybersecurity and biological domains. Coordinated response capabilities will be enabled through cross-sector collaboration mechanisms facilitating rapid information sharing during incidents. Data protection will be strengthened through biology-appropriate controls designed specifically for genomic information and other sensitive biological data. Additionally, vulnerability to supply chain attacks should be reduced through enhanced vendor oversight and secure procurement processes for biological materials and services.

However, significant limitations exist. Implementation costs may be prohibitive for smaller organizations with limited cybersecurity budgets. Framework effectiveness depends on cross-sector cooperation that may prove difficult to achieve given competitive pressures and information sharing concerns in commercial biotechnology. The rapidly evolving threat landscape requires continuous framework adaptation, demanding ongoing investment and attention. Lack of empirical validation means actual performance metrics remain unknown until real-world testing occurs.

Success requires sustained organizational commitment and resource allocation over multiple years, which may be challenging to maintain through leadership changes and competing priorities. The framework represents a conceptual architecture requiring real-world testing, refinement, and validation. Organizations should adapt components to their specific contexts rather than attempting wholesale adoption without customization.

8. Conclusion

This literature synthesis establishes cyberbiosecurity as a critical emerging challenge requiring immediate attention across biotechnology, healthcare, and defense sectors. Our review demonstrates that traditional cybersecurity frameworks are insufficient for cyber-biological threats, particularly given documented incidents such as healthcare ransomware attacks (Landi, 2019; National Audit Office, 2018) and proof-of-concept research showing DNA-encoded malware as feasible attack vectors (Ney et al., 2017).

The proposed Integrated Cyberbiosecurity Defense Framework (ICDF) addresses these challenges through six components: Bio-Risk Assessment, Advanced Threat Detection, Simulation Platforms, Cross-Disciplinary Collaboration Centers, Enhanced Data Protection, and Workforce Development. This framework represents the first comprehensive architecture specifically tailored to biological systems and cyber-biological convergence.

Critical next steps include empirical validation through pilot implementations in healthcare, biotechnology, and research organizations to test the framework's practical applicability. The development of standardized cyberbiosecurity metrics is essential to enable consistent measurement and comparison of security postures across organizations. Establishment of regulatory guidance for cyber-biological risks would provide organizations with clear compliance requirements and reduce the current uncertainty around legal obligations. Investment in interdisciplinary workforce development programs is necessary to build the pipeline of professionals with combined expertise in cybersecurity and life sciences. Finally, creation of international cooperation mechanisms would enable coordinated threat intelligence sharing and response capabilities across national borders, recognizing that cyber-biological threats do not respect geographic boundaries. While implementation costs will exceed traditional cybersecurity programs, the documented impact of cyber incidents in healthcare and biotechnology - including operational disruption, intellectual property loss, and potential public health consequences - demonstrates the necessity of specialized defensive measures.

The convergence of AI, synthetic biology, and advancing cyber capabilities (Hu et al., 2020; Zhang et al., 2020; Yoo, 2020) creates unprecedented vulnerabilities extending beyond data breaches to potential manipulation of biological systems themselves. Regulatory frameworks (Office of Science and Technology Policy, 2022), educational programs, and international cooperation must evolve to meet this challenge.

Future research priorities include controlled testing of proposed framework components to validate their effectiveness in real-world settings. Longitudinal studies of threat evolution are needed to track how cyber-biological risks change over time and ensure the framework remains relevant. Researchers should develop validated effectiveness metrics that can objectively measure cyberbiosecurity improvements across different organizational contexts. Economic analysis comparing implementation costs against incident prevention benefits would help organizations justify investments in cyberbiosecurity programs. Finally, investigation of international governance mechanisms is essential to address the transnational nature of cyber-biological threats and enable coordinated global response capabilities. This paper provides a foundation for systematic cyberbiosecurity development, but practical validation and continuous adaptation remain essential. The biological domain's criticality to public health, food security, and economic stability makes effective cyberbiosecurity frameworks an urgent priority requiring sustained investment and cross-sector commitment.

Ethics Declarations: This work relies exclusively on published, publicly available literature, government reports, and audited incident investigations; no new human-subjects research, surveys, or interviews were conducted for this version of the analysis. Limitations include potential publication bias in incident reporting, the evolving nature of cyberbiosecurity threats (new technical capabilities and incidents may arise), and restricted availability of classified or private attribution details. Future work should complement the literature synthesis with structured primary research (surveys, interviews, and technical red-team exercises) if IRB approval and participant protections are obtained.

AI declaration: Artificial Intelligence tools were utilized in limited capacity during the development of this research paper for literature organization, reference formatting consistency, and editing support for clarity and professional presentation of complex technical concepts. The Integrated Cyberbiosecurity Defense Framework (ICDF) and its components represent novel conceptual contributions developed by the research team through systematic literature synthesis. All framework design, thematic analysis, and conclusions are the original intellectual work of the authors. AI tools were not used for literature selection, analysis interpretation, or generation of research conclusions. All content was verified for accuracy and originality by the authors.

References

Berger, M. (2021) 'The dark web and biotechnology risks', Nature Biotechnology, 39(4), pp.396–397. doi:10.1038/s41587-021-00896-8.

Brumfiel, G. (2020) 'DNA sequencing devices can be hacked', Nature, 579(7797), pp.16–17. doi:10.1038/d41586-020-00694-9.

Choucri, N., Madnick, S. and Ferwerda, J. (2019) 'Insiders and cybersecurity: Threats and countermeasures', Communications of the ACM, 62(2), pp.26–28. doi:10.1145/3303867.

ENISA (2021) Threat landscape for healthcare. Heraklion: European Union Agency for Cybersecurity. Available at: https://www.enisa.europa.eu/publications (Accessed: 2 October 2025).

Goodman, B. (2020) ‘Ethical challenges of digital health’, BMC Medical Ethics, 21(1), p.42. doi:10.1186/s12910-020-00478-4.
Hardin, J. (2021) ‘Securing bioinformatics pipelines: Threat modeling approaches’, Bioinformatics, 37(11), pp.1624–1631. doi:10.1093/bioinformatics/btaa883.
Hu, J., Gao, Y. and Liu, J. (2020) ‘Potential cybersecurity threats in synthetic biology’, Nature Communications, 11(1), p.3384. doi:10.1038/s41467-020-17102-9.
Kim, J. (2021) ‘Cyber resilience in healthcare IoT networks’, IEEE Internet of Things Journal, 8(12), pp.9512–9521. doi:10.1109/JIOT.2020.3048238.
Kruse, C.S., Frederick, B., Jacobson, T. and Monticone, D.K. (2017) ‘Cybersecurity in healthcare: A systematic review of modern threats and trends’, Technology and Health Care, 25(1), pp.1–10. doi:10.3233/THC-161263.
Landi, H. (2019) ‘Ransomware attacks in healthcare sector surged in 2019’, Healthcare IT News. Available at: https://www.healthcareitnews.com/news/ransomware-attacks-healthcare-sector-surged-2019 (Accessed: 2 October 2025).
Li, X., Zhang, Y. and Sun, Y. (2021) ‘Cyber-physical attacks and defenses in medical systems’, IEEE Transactions on Cybernetics, 51(9), pp.4564–4575. doi:10.1109/TCYB.2020.2985166.
Liu, L. (2019) ‘The future of genomic data privacy’, Trends in Genetics, 35(7), pp.429–431. doi:10.1016/j.tig.2019.04.001.
Martin, G., Ghafur, S., Kinross, J., Hankin, C., Darzi, A. and Jalil, R. (2017) ‘WannaCry: A year on’, BMJ, 361, k2381. doi:10.1136/bmj.k2381.
Miller, T. (2020) ‘Security issues in biomedical engineering’, Annual Review of Biomedical Engineering, 22, pp.459–483. doi:10.1146/annurev-bioeng-092019-030838.
Murch, R.S., So, W.K., Buchholz, W.G., Raman, S. and Peccoud, J. (2018) ‘Cyberbiosecurity: An emerging new discipline to help safeguard the bioeconomy’, Frontiers in Bioengineering and Biotechnology, 6, p.39. doi:10.3389/fbioe.2018.00039.
National Academies of Sciences, Engineering, and Medicine (2020) Safeguarding the bioeconomy. Washington, DC: National Academies Press. doi:10.17226/25525.
National Audit Office (2018) Investigation: WannaCry cyber attack and the NHS. London: NAO. Available at: https://www.nao.org.uk/report/investigation-wannacry-cyber-attack-and-the-nhs/ (Accessed: 2 October 2025).
National Institute of Standards and Technology (2020) Framework for improving critical infrastructure cybersecurity: Version 1.1 (NIST Special Publication 800-53 Rev. 5). Gaithersburg: NIST.
Ney, P., Koscher, K., Organick, L., Ceze, L. and Kohno, T. (2017) 'Computer security, privacy, and DNA sequencing: Compromising computers with synthesized DNA, privacy leaks, and more', in Proceedings of the 26th USENIX Security Symposium. Vancouver: USENIX Association, pp. 765–779.
Nix, R., Zhang, S., Tan, J. and Peccoud, J. (2021) ‘Cybersecurity risks in laboratory automation’, Nature Biotechnology, 39(8), pp.946–947. doi:10.1038/s41587-021-00970-1.
Office of Science and Technology Policy (2022) National security memorandum on strengthening biosecurity. Washington, DC: OSTP.
Peccoud, J. (2018) ‘Cyberbiosecurity: From naive trust to risk awareness’, Trends in Biotechnology, 36(1), pp.4–7. doi:10.1016/j.tibtech.2017.10.012.
Peccoud, J. and Peccoud, S. (2020) ‘Preparing for cyberbiosecurity: A policy framework’, Biotechnology Advances, 42, p.107592. doi:10.1016/j.biotechadv.2019.107592.
Richardson, L.C., Murch, R.S. and Peccoud, J. (2019) ‘Cyberbiosecurity: A call for cooperation in a new threat landscape’, Frontiers in Bioengineering and Biotechnology, 7, p.99. doi:10.3389/fbioe.2019.00099.
Smith, H.J. (2018) ‘Cybersecurity and data breaches: A growing public health concern’, American Journal of Public Health, 108(8), pp.989–990. doi:10.2105/AJPH.2018.304528.
U.S. Department of Health and Human Services (2018) Health industry cybersecurity practices: Managing threats and protecting patients (HICP Publication 405-D-0258). Washington, DC: HHS.
U.S. Department of Homeland Security (2019) Cyberbiosecurity: Next steps in securing the bioeconomy (DHS Publication 2019-CISA-01). Washington, DC: DHS.
Weiss, S. (2019) ‘Protecting biological data: Security measures in genome research’, Journal of Genomics and Informatics, 17(1), p.e7. doi:10.5808/GI.2019.17.1.e7.
Yoo, S. (2020) ‘Cyber threats in digital health innovation’, JMIR Medical Informatics, 8(2), p.e16725. doi:10.2196/16725.
Zhang, B. and Wang, H. (2020) ‘DNA data storage and its security challenges’, Frontiers in Bioengineering and Biotechnology, 8, p.581407. doi:10.3389/fbioe.2020.581407.
Zhang, Y., Deng, R.H. and Liu, Y. (2020) ‘Security and privacy in smart healthcare: Research challenges’, IEEE Internet of Things Journal, 7(5), pp.4424–4448. doi:10.1109/JIOT.2019.2942190.

Cybersecurity Issues and Solutions Within the South African Small and Medium-Sized Enterprises

Benediction Kitwa Kalombola and Tabisa Ncubukezi
Cape Peninsula University of Technology, Cape Town, South Africa

222689900@mycput.ac.za
ncubukezit@cput.ac.za

Abstract: This systematic literature review paper presents the main cybersecurity issues and their challenges within the Small and Medium-sized Enterprises (SMEs) in South Africa. The SME's play a very crucial role in the growth of the country's economy and Gross Domestic Product (GDP). However, following the global pandemic, the economy has become increasingly dependent on technology as a driver in different sectors. With the rapid adoption of technology, SMEs constantly face significant cybersecurity issues and challenges that regularly demand proactive and effective measures. These measures guard against exposure to sophisticated cyber-attacks. This study examines the state of cybersecurity within South African SMEs by assessing the effective use of cybersecurity measures, exploring the extent of their implementation, and identifying best practices to strengthen cyber resilience within these SMEs. To achieve this aim, a systematic literature review is conducted to examine high-quality peer-reviewed journals from 2021 to July 2025, providing more insight into the issues associated with cybersecurity in SMEs. Results highlight the vulnerability of SMEs to various cyber threats, including phishing, insider threats, and ransomware. The lack of awareness among employees, inadequate cybersecurity measures, limited resources, the shortage of professional experts, and the absence of effective measures specifically tailored to the needs of SMEs. Additionally, the mitigation strategies emphasize the adoption of robust security measures that are specifically tailored to the needs of SMEs. To enhance cyber resilience, this study also recommends the use of cybersecurity measures tailored to SMEs, encouraging employee education training programs, and the creation of reliable strategies.

Keywords: Best practices, Cybersecurity, Cybersecurity challenges, Cybersecurity issues, Cybersecurity measures, Cyber threats, Small businesses, Systematic literature review

1. Introduction

In South Africa, small and medium-sized enterprises (SMEs) play a crucial role in driving economic growth; however, SMEs have become increasingly vulnerable to cyber threats with the rapid adoption of technology. Despite the benefits associated with digital transformation in improving operational activities, it has also introduced new threats, leaving SMEs exposed to threats such as data breaches, ransomware, and phishing, mainly due to the lack of awareness, limited resources, and ineffective measures (Fakir, 2024; Saeed *et al.*, 2023; Ncubukezi, 2022a). Additionally, because SMEs often rely on outdated security measures and ineffective strategies, they are frequently left vulnerable to threats, particularly in the financial and healthcare sectors (Mahlangu, 2023; Ncubukezi, 2022b).

SMEs encounter issues and challenges in adopting effective cybersecurity measures due to inadequate management and regulations, the absence of skilled professionals, and resources (Shingange, 2022). According to Phahlamohlaka *et al.*, (2022), there is a gap between the practical implementations and available frameworks, including the Zero Trust Architecture (ZTA), National Cybersecurity Policy Framework (NCPF) and the National Institute of Standards and Technology (NIST), which is further enlarged by SMEs unawareness and inability to adopt advanced tools such as multi-factor authentication, encryption, and artificial intelligence (AI) and machine learning (ML) for early threat detection (Admass *et al.*, 2024). Considering that SMEs employ almost 60% of the working population and constitute around 90% of registered businesses in the nation, the impact of cyber threats on the financial institutions is enormous in South Africa. Because human mistakes are the primary reasons that trigger data breaches, Reshmi (2021) argues that this is largely due to the lack of adequate cybersecurity training among employees.

Cybersecurity has become a critical concern for South African SMEs. Businesses are increasingly exposed to sophisticated threats and face challenges in protecting their data (Ncubukezi & Mwansa, 2021). As technology adoption advances, it has become increasingly crucial for SMEs to comprehend the complexities of cybersecurity threats in order to survive and thrive (Benjamin *et al.*, 2024). The problem stems from the lack of effective security measures and the reliance on outdated systems, which are now less effective in mitigating sophisticated threats. The emergence of advanced cyber threats causes several damages within SMEs, including operational disruption, loss of reputation, and economic loss, making it difficult for these businesses to protect the sensitive data they handle (Dave *et al.*, 2023; Mahlangu, 2023).

While several studies addressed the key importance of affordable and customized security measures in which SMEs' real-world challenges are considered, there is still a gap regarding the specific issues and challenges encountered by SMEs along with practical and tailored solutions (Benjamin *et al.*, 2024). The lack of modern and tailored security measures leaves the SMEs vulnerable and constantly exposed to sophisticated cyber threats. Without addressing this, both SMEs and financial institutions will remain vulnerable to cyber threats and suffer the associated consequences. Hence, the study intends to examine the major issues and challenges, exploring the extent of implementation of existing security measures, and providing best practices and effective measures to enhance cyber resilience within SMEs.

To achieve the aim, this work will carry out the following objectives:

- To investigate the major cybersecurity issues experienced by the SMEs in South Africa.
- To explore the extent of implementation of existing measures.
- To investigate the best practices for effective security measures.

2. Literature Review

Several studies have examined the ever-evolving field of cybersecurity. Despite this, a thorough examination of this study is imperative, considering the specific obstacles that most Small and Medium-sized Enterprises (SMEs) face in South Africa (Fakir, 2024). While studies provide recommendations about security threats, the challenges experienced by SMEs remain largely unaddressed (Benjamin *et al.*, 2024). Hence, there is a need for local studies that strongly focus on the specific challenges faced by these businesses. SMEs are becoming more vulnerable to sophisticated cyber threats. However, studies have revealed the challenges these businesses frequently experience in adopting effective security measures, highlighting the need for more advanced strategies tailored to them (Saeed *et al.*, 2023). However, there is still a gap between awareness, current strategies and the actual implementation and regulations (Phahlamohlaka *et al*., 2022). The gradual growth of the technology-driven economy has brought both benefits and emerging threats to SMEs, despite their significant contribution to the economy. Due to digital transformation, these businesses face several issues and challenges in protecting their digital assets, primarily due to a lack of effective security measures and resource constraints (Fakir, 2024), leaving them vulnerable and defenceless against sophisticated threats (Dave *et al.*, 2023). While more advanced frameworks, such as NIST and Zero Trust Architecture (ZTA), exist, their adoption remains limited due to the lack of skilled professionals, financial support, and proper policies required for their implementation (Bokan & Santos, 2021; Lokare *et al.*, 2025). Despite growing awareness of cybersecurity risks, a gap remains between knowledge, strategies, and actual practical actions, especially in resource-constrained environments, including SMEs (Lokare et al., 2025; Admass et al., 2024).

2.1 Summary of Existing Literature

Benjamin et al. (2024) highlight that SMEs are the primary target of cyber threats due to the lack of effective measures and skilled professionals while adopting technology. SMEs are frequently exposed to cyber threats due to inadequate security measures, underscoring the need for more comprehensive and tailored security measures to enhance the protection of their digital data and ensure smooth business operations. Mahlangu (2023) examines the growing concern that emerging ransomware attacks have become a critical concern for financial institutions. The leadership, as well as the behaviour of people, plays a vital role in strengthening endurance within organizations (Mahlangu, 2023), drawing attention to the fact that defense against ransomware is unlikely to be effectively eradicated with technical implementations only. In consideration of risks and challenges faced by financial institutions and businesses, Mahlangu (2023) advocates for measures specifically tailored to these institutions (Mahlangu, 2023). Admass et al. (2024) investigate the role of Machine Learning (ML) and Artificial Intelligence (AI) to strengthen cybersecurity. Admass et al. (2024) highlight the difficulties resulting from insufficient knowledge and flexibility commonly encountered by the majority of SMEs, despite the fact that potential solutions could arise from technological advances such as these. In most cases, SMEs are unable to guarantee effective security due to universal security measures that fail to meet the specific needs of SMEs (Admass et al., 2024). Hence, there is an urgent need for specifically tailored measures to address the challenges faced by SMEs. The study by Saeed *et al.* (2023) highlights that while the adoption of technology may contribute to business growth, it also exposes frequent SMEs to cyber threats. In addition to fostering risk-based security strategies, Saeed *et al.* (2023) underline a pressing need for proactive measures tailored to SMEs, while also suggesting the adoption of flexible and innovative strategies. The study by Fakir (2024) examines the most significant challenges associated with adopting cybersecurity. Because almost every organization operates with various activities and management structures, the difficulties

also vary from one to another organization (Fakir, 2024). Hence, there is a need for more adaptive approaches tailored to the specific needs and challenges faced by SMEs, due to the ineffectiveness of the majority of generic solutions. Additionally, there is a need for continued investigations to closely monitor the advancement of cybersecurity requirements within SMEs (Fakir, 2024).

Shingange (2022) focuses on the lack of connection that exists between the design of measures and their implementation in practice, while investigating the challenges associated with cybersecurity laws and regulations in South Africa. Shingange (2022) highlights the ongoing absence of communication between private industries and governments, which weakens the overall effectiveness of cybersecurity operations. Due to a failure to enforce current laws and regulations, there is a need to critically evaluate and strengthen them (Shingange, 2022).To ensure successful progression, there is a need for greater coordination along with further visible enforcement. In 2021, Reshmi investigated ransomware attacks and their prevention mechanisms. While more advanced technologies like Artificial Intelligence (AI) and Machine Learning (ML) have the potential to identify threats early, their application remains limited within SMEs. This study suggests a wider integration of AI/ML to enhance security against ransomware.

The CENSOR model, as outlined by Tsiodra et al. (2023), is a risk-based model for cybersecurity decision-making. It is specifically beneficial to the SMEs as it guides them in prioritizing their limited resources by evaluating threats according to their significance. The authors also provide an illustration demonstrating how SMEs can apply this model to allocate resources more effectively. Tsiodra et al. (2023) explore the relationship between the magnitude of cyber threats that SMEs encounter and the level of investment that SMEs make. Hence, the critical importance of establishing an appropriate balance between budgeting and maintaining effective security. In the same way, Hermanus (2023) examined the vulnerabilities associated with information security within the financial environment in Cape Town. As reported by the participants in the interviews, organizations continue to face serious problems towards the full adoption such as insufficient resources, the shortage of professional experts, and a poor organized response strategy, despite the fact that some enterprises started putting into effect security awareness training, regulations to the protection of information, intrusion detection systems as well as authentication and authorization systems (Hermanus (2023). Hence, this reveals that some enterprises are becoming aware and starting to take effective action; however, the complete adoption is constrained by operational issues.

In 2023, Musa investigated the cybersecurity issues faced by SMEs from a systems engineering perspective. This study highlights the lack of ongoing monitoring, employee training, and incident response, as the majority of SMEs rely on basic strategies such as antivirus, and password policies. The absence of skilled professionals, resource shortages, and the need for tailored security measures are a few barriers preventing the adoption of cybersecurity. In their review of literature, Alveera & Sukira (2024) delt with cybersecurity within the SMEs, highlighting their susceptibility to cyber-attacks from malicious computer programs and ransomware, regardless of the importance of these SMEs to the growth of the economy within the country, which is mainly the consequence of several factors, including the limited financial resources, inadequate or outdated infrastructures, and regular misperception of the complexity of cybersecurity. Additionally, Alveera & Sukira (2024) highlight the fact that SMEs are often given little attention compared to large corporations, which are mainly prioritized by government initiatives for cybersecurity. This study recommends the adoption of rigorous security mechanisms specifically tailored to the needs of SMEs, conducting risk assessments, and educating team members about cyber threats to ensure robust cybersecurity for SMEs (Alveera & Sukira, 2024).

In 2024, Mugwagwa *et al.* highlight that due to poor awareness, resource constraints, and the absence of skilled professionals, SMEs remain vulnerable to complex threats like insider threats, ransomware, and phishing, highlighting the need for effective malware detection, multi-factor authentication, encryption, along regular employee education to effectively enhance cyber resilience within SMEs. In summary, while there is a growing awareness of cybersecurity, a significant gap remains between awareness, strategic plans, and actual implementation. While cutting-edge technologies like AI/ML can enhance cyber resilience within SMEs, their adoption remains limited due to financial constraints, resource shortages, and a lack of skilled professionals, which are primary issues faced by SMEs in South Africa, preventing them from implementing effective cybersecurity measures.

3. Methodology

A comprehensive systematic literature review is employed in this study to investigate the state of cybersecurity in South African SMEs, with a strong focus on the major issues and challenges, as well as the

extent of implementation of existing security measures. The objective is to provide best practices for enhancing their resilience to cybersecurity. The systematic literature review is adopted in this study along with the analysis of the contents to discover, critically review, and prioritize key research that directly aligns with the previously established questions (using a systematic literature review) and come up with commonly identified issues and challenges, as well as important and repeated solutions and recommendations (content analysis).

3.1 Selection Criteria

To ensure a comprehensive examination of the available literature, the data used were collected from various important databases, including Google Scholar, IEEE Xplore, ScienceDirect, and MDPI. Additionally, South Africa's sources were particularly included to enable a more objective and in-depth examination of the challenges associated with cybersecurity in the South African SMEs.

3.2 Search Strategy

To gather relevant sources of information, a systematic search was conducted. The search was conducted using keywords and phrases including "Cybersecurity" AND "South Africa" AND "SMEs," “Cybersecurity in SMEs,” “Cybersecurity,” “Cybersecurity challenges,” “Information security in SMEs,” Online security,” “Cybersecurity issues,” “Digital security,” “Cyber issues,” “Cyber resilience strategies for SMEs,” and “Data breaches.”

3.3 Prisma Flow Diagram

This study adheres to the PRISMA principle to ensure clarity in the selection of literature, with the aim of identifying, examining, and selecting the most suitable studies for this research. In accordance with predefined inclusion and exclusion criteria, the overall approach consisted of several narrowing phases. For this reason, the following Figure 1 below presents an illustration of the method by which information progresses throughout the different stages of this process of review:

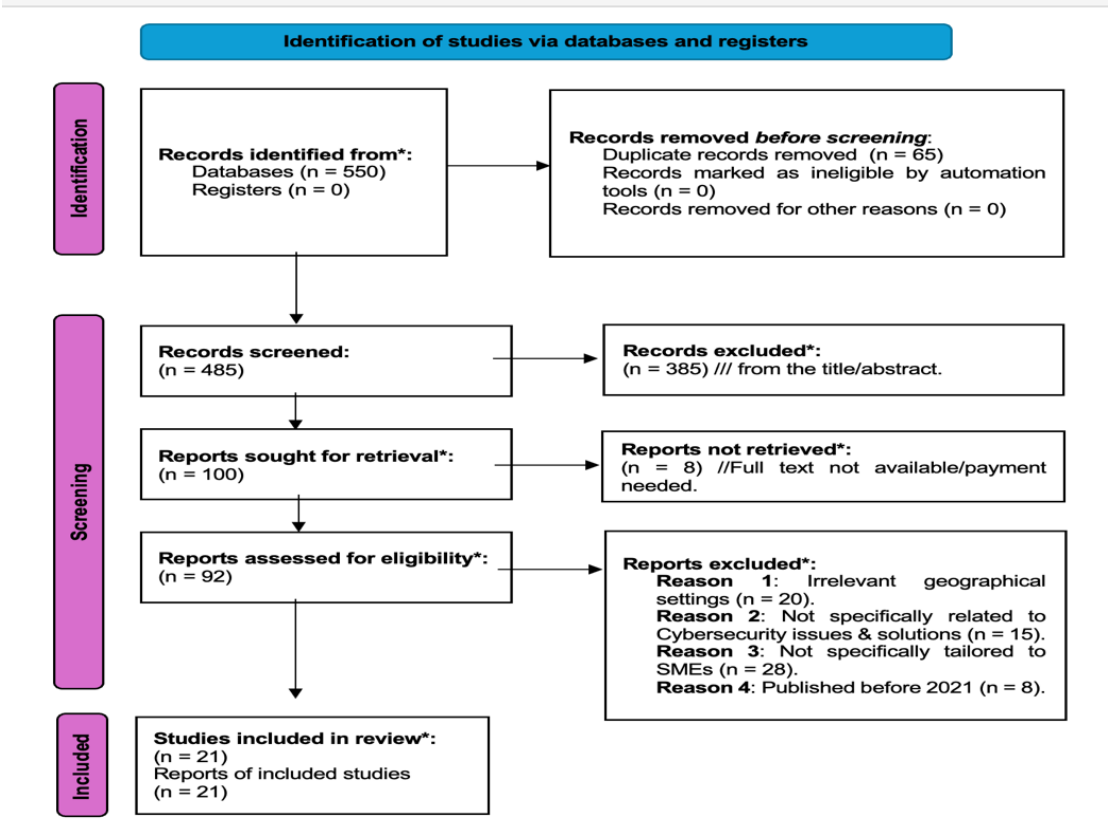


Figure 1: Prisma Flow Diagram

3.4 Inclusion and Exclusion Criteria for Relevant Literature

To gather more accurate and relevant studies, the inclusion and exclusion criteria were defined in this study. Therefore, this research examined various studies that discuss cybersecurity issues and challenges within SMEs, with a focus on studies that emphasize the development of security measures specifically tailored to the needs of SMEs. Additionally, only English articles published between year 2021 and July 2025 were considered to collect the most recent information. On the other hand, excluded studies were those that did not focus on SMEs' issues, challenges, and their needs. Additionally, articles published before 2021 were excluded to ensure that this investigation aligns successfully with the latest information and developments. By following these guidelines, the review was focused on the predefined objectives.

3.5 Data Analysis

To identify important themes, challenges, and technological innovations associated with cybersecurity for SMEs, the content analysis was conducted on the selected literature, aiming to identify issues and challenges encountered by SMEs and potential innovative solutions to enhance their cyber resilience. Finally, the results were combined to provide actionable recommendations to enhance cyber resilience within South African SMEs.

4. Results and Discussions

A total of four interrelated points were revealed from the review of the literature: the different kinds of threats commonly associated with cybersecurity that SMEs experience, the issues preventing them from adopting and using effective security measures, and the extent to which these measures are successfully implemented. SMEs have begun to understand the necessity of introducing effective cybersecurity measures, as the digital economy resulting from technological advancements continues to evolve. The use of these effective cybersecurity measures has become crucial for ensuring the protection of the digital assets they contain while maintaining business operations. The goal of this systematic literature review was to explore the primary issues, challenges, and potential solutions related to the effective implementation of cybersecurity in South African SMEs. Categorized into four key themes, the results of this research indicate that while threats to cybersecurity are gradually being recognized, the potential to effectively establish robust and effective countermeasures remains remarkable. In this section, the findings will be discussed along with their connection to the research objectives and questions.

4.1 Common Cybersecurity Threats Experienced by SMEs in South Africa

SMEs remain vulnerable to various cyber threats identified in the literature, including poor password policies, ransomware attacks, phishing, and insider threats (Mugwagwa et al., 2024; Reshmi, 2021). For businesses and financial institutions, ransomware poses a significant concern due to its potential to result in financial loss and business disruption (Mahlangu, 2023). SMEs have become vulnerable to various threats because of digital transformation, despite the benefits it has brought. Despite the growing awareness of cybersecurity among SMEs, the results of the study indicate that they frequently experience various issues and challenges that restrict them from effectively making use of all necessary measures, including the lack of skilled professionals, poorly structured organizations, outdated systems, and budget constraints (Hermanus, 2023; Musa, 2023; Benjamin *et al*., 2024). Issues with policies and regulations are also quite important. According to Shingange (2022), an inefficient approach to cybersecurity governance has resulted from the lack of strong enforcement and inadequate interaction between the public and commercial sectors. It is common for many SMEs to operate without clearly established cybersecurity policies and regulations, instead relying on improvised solutions, which highlights that cybersecurity is a problem that extends beyond technology and has a profound impact on the way organizations and their systems interact. Given this, there is a need for robust organizational support, effective leadership, and thoughtful policy interventions.

Additionally, Tetteh (2024) emphasizes that more structured threat detection, combined with incident reporting, is essential for achieving more effective cybersecurity in SMEs. According to Tetteh (2024), cyber resilience across the nation has become weaker due to the inability of SMEs to report security breaches, which limits the opportunities for education across the industry. Therefore, there is a need for simplified frameworks that allow SMEs to effectively report incidents and assess risks in a more adaptive manner.

4.2 Challenges Associated with the Effective Use of Cybersecurity Measures

The gap between knowledge and practical actions remains large. The reliance on obsolete systems, financial constraints, a lack of skilled professionals, and the need for tailored security measures are the primary

obstacles (Hermanus, 2023; Musa, 2023; Benjamin et al., 2024). Moreover, further challenges include the breakdown of collaboration between the private and public sectors, inconsistent directions for implementation, and a failure to enforce regulations, as well as inadequate management and weak internal processes (Shingange, 2022; Fakir, 2024). Additionally, the failure of SMEs to report cyber incidents significantly impacts the general and national response efforts. This highlights an ongoing issue with inadequate reporting and the lack of a structured prioritization framework (Tetteh, 2024). By utilising a practical model introduced in this study, Small and medium-sized enterprises are enabled to focus their resources on addressing the most urgent issues and challenges, which in turn enhances their overall resilience.

Although small and medium-sized enterprises (SMEs) are becoming aware of cybersecurity risks and challenges they experience, the review shows that the efforts they make to put in place practical and effective preventive measures are still at the basic level, which include antivirus software, constrained awareness education, and password regulations, while more advanced and sophisticated tools such as ongoing monitoring, systems for intrusion detection, and the Zero Trust Architecture (ZTA) remain rarely, or not used (Hermanus, 2023; Musa, 2023; Erdogan *et al*., 2023; Lokare *et al*., 2025). This circumstance illustrates a clear disconnect between awareness and action. Small and medium-sized enterprises (SMEs) often struggle to acquire cutting-edge technologies, primarily due to financial constraints and limited resources. Additionally, their ability to maintain or enhance complicated systems is limited due to their lack of technical expertise. In many cases, the use of cybersecurity is generally more of a response to specific incidents than an anticipatory measure, as highlighted in section 4.2 "Long-term Strategy," where businesses commonly wait for incidents to occur before taking action.

4.3 Extent of Implementation of Cybersecurity Measures

According to the reviewed articles, many SMEs are using basic security precautions, ranging from password rules, limited awareness training for employees, and anti-virus programs, while preventive and more sophisticated strategies exist, but continue to be left behind (Hermanus, 2023; Musa, 2023; Erdogan et al., 2023; Lokare et al., 2025). Additionally, certain organizations possess disaster recovery plans, but these plans often lack clear procedures. Due to their high costs and complexity, emerging technologies such as threat identification based on intelligence and machine learning (AI/ML), intrusion detection systems, and ongoing monitoring are rarely adopted (Admass et al., 2024; Lokare et al., 2025).

A variety of effective approaches for enhancing SMEs' resilience to cybersecurity threats are highlighted in this research, pointing out the adoption of risk-based strategies to enable SMEs to focus their resources on the greatest threats (Tsiodra *et al.*, 2023; Tetteh, 2024) and the necessity for international models, including the NIST framework for cybersecurity and the ZTA, to be adjusted for resource-constrained environments (Fakir, 2024; Bokan & Santos, 2021). Moreover, studies underlining the necessity for ongoing staff education and training which is indispensable for preventing errors among employees (Mugwagwa *et al.*, 2024; Benjamin *et al.*, 2024) while also promoting encryption, multi-factor authentication, and the need to identify threats at their early stage by integrating more advanced technologies such as AI/ML (Admass *et al.*, 2024; Reshmi, 2021). Additionally, studies promote collaboration between the private and public sectors to offer SMEs adequate funding and support (Shingange, 2022).

4.3.1 Best practices and recommendations

Studies suggest that to reconcile the divide between knowledge and practical action, it is necessary to tailor cybersecurity approaches to the needs of SMEs, adapting frameworks like Zero Trust Architecture (ZTA) and NIST to resource-constrained environments (Fakir, 2024; Bokan & Santos, 2021). To reduce human errors, a continuous educational program becomes crucial (Mugwagwa et al., 2024; Benjamin et al., 2024), as well as risk-based approaches for effective resource allocation (Tsiodra et al., 2023). Moreover, advanced measures such as AI/ML, multi-factor authentication, and encryption can enhance security through early threat detection (Admass et al., 2024; Reshmi, 2021). Additionally, studies suggest that the government should strengthen collaboration between the private and public sectors, increase policy monitoring, and provide more detailed recommendations tailored to the benefit of SMEs.

The need for more effective and robust cybersecurity measures has become paramount, particularly within SMEs, which continue to remain more vulnerable and are constantly exposed to cyber threats as a result of being primary targets. Thus, the following recommendations resulting from the results and conclusions drawn from this research will enable SMEs to further strengthen their resilience and utilization of cybersecurity in South Africa:

- **Promote awareness on cybersecurity**: To effectively handle human-related vulnerabilities while improving the culture of companies, there is a need for regular awareness and educational campaigns about cybersecurity.
- **Adapt frameworks tailored to SMEs' needs:** A necessity for both policymakers as well as experts in cybersecurity to design more reasonable security measures that adapt to the specific business processes, size, and the available resources of SMEs.
- **Promote the utilization of risk-based solutions**: As recommended by Tetteh (2024) and Tsiodra *et al.* (2023), SMEs could first detect and address high-risk spots, concentrating the insufficient resources they have on the biggest threats to cybersecurity.
- **Foster cooperation**: To ensure long-term viability within the SMEs, the findings of this research suggest collaboration between the private sector, government agencies, and industries in order to offer more affordable accessibility to support, financial resources, and equipment.
- **Promote the reporting of security incidents**: To gain more insight from cyber disasters and effectively strengthen cyber resilience within the nation, this study recommends the start-up of more secure and accessible ways to report incidents.

Table 1: Summary of cybersecurity threats, challenges, and practices in SMEs

Theme	Key Findings
Common cybersecurity threats	In S.A, SMEs are constantly confronted with a range of threats such as ransomware, malware, data breaches, and phishing, which frequently arise from weak credentials, lack of proper monitoring, and outdated systems.
Challenges to implementation	Primary obstacles consist of the lack of skilled cybersecurity professionals, poor enforcement of regulations, a general lack of awareness, lack of financial resources, and the struggle to convert national policies (e.g. NCPF) into effective action.
Extent of cybersecurity implementation	Despite the growing awareness about the importance of cybersecurity, implementation remains low. SMEs frequently rely on basic security measures such as Antivirus and firewalls, while neglecting more advanced strategies such as AI-based threat detection, well-structured risk management frameworks, and the MFA.
Best practices and recommendations	To enhance cybersecurity, it is essential to engage in ongoing training, refining password policies, using budget-friendly frameworks, using the CENSOR for resource allocation, promoting cooperation between governments, SMEs, and private sectors cybersecurity providers, and setting priorities based on risks.

5. Conclusion and Future Work

This study aimed to investigate issues and challenges associated with cybersecurity experienced by SMEs in South Africa, exploring the extent of implementation of existing security measures and providing potential best practices. The research reveals a poor state of cybersecurity within SMEs regardless of the growing awareness of cybersecurity threats. These businesses face ongoing challenges in adopting cutting-edge technological solutions due to a lack of employee awareness, insufficient resources, and a shortage of skilled professionals. Moreover, the practical integration of advanced security frameworks, such as ZTA and NIST, remains limited due to their high implementation cost, despite the potential they offer, which explains the reliance of SMEs on basic security strategies, including antivirus software and password policies. Additionally, the results suggest that SMEs often opt for a reactive plan of action, prioritizing a response to incidents over proactive preparation. However, cybersecurity should be considered both a strategic and an administrative issue, rather than a technical one only, as it is an essential way to handle all these issues. For South African SMEs to enhance their cyber resilience, it is necessary to integrate cybersecurity into their organizational culture. For this reason, the ultimate necessity is for security measures that are more affordable, more tailored to the specific needs of SMEs, capable of accommodating issues associated with their activities and resource constraints, thus strengthening the resilience of SMEs in South Africa to cybersecurity threats.

For future work, - further studies could be conducted on the improvement of small and medium-sized enterprises' (SMEs) potential towards threat detection and response in real-time, by incorporating algorithms as well as data structures within a programmed information security system, which could provide more

understanding on how the usage of these approaches could be improved to be beneficial in environments with restricted resources.

AI declaration: For the development of this work, no AI tools were used.

Ethics declaration: Since the study adopts a systematic review of previous literatures which focus on secondary data publicly made available without direct human interaction or raw data collection, no ethical approval was required. However, in order to preserve academic integrity and avoid plagiarism, all sources have properly been referenced.

References

Admass, W., Munaye, Y. & Diro, A. 2024. Cyber security: State of the art, challenges and future directions. Amsterdam: Elsevier BV.

Alveera, T. & Sukira, M. 2024. Cybersecurity challenges and resilience strategies in South African SMEs. Cape Town: UCT Press.

Benjamin, L., Adegbola, A., Amajuoyi, P., Adegbola, M. & Adeusi, K. 2024. Digital transformation in SMEs: Identifying cybersecurity risks and developing effective mitigation strategies. Vol. 19. GSC.

Bokan, B. & Santos, J. 2021. Managing cybersecurity risk using threat-based methodology for evaluation of cybersecurity architectures. New York: IEEE.

Dave, D., Sawhney, G., Aggarwal, P., Silswal, N. & Khut, D. 2023. The new frontier of cybersecurity: Emerging threats and innovations. Toba: IEEE.

Erdogan, T., Rahman, M. & Kwon, H. 2023. Assessing cybersecurity maturity among small and medium-sized enterprises. New York: IEEE.

Fakir, J. 2024. An investigation of cybersecurity implementation challenges among South African SMEs. Master's dissertation. Johannesburg: University of the Witwatersrand.

Hermanus, K. 2023. Information security vulnerabilities in financial institutions in Cape Town. Master's dissertation. Cape Town: University of Cape Town.

Lokare, A., Bankar, S. & Mhaske, P. 2025. Integrating cybersecurity frameworks into IT security: A comprehensive analysis of threat mitigation strategies and adaptive technologies. arXiv preprint. Available at: https://doi.org/10.48550/arXiv.2502.00651. (Accessed : 15 August 2025).

Mahlangu, N. 2023. Strategies to mitigate ransomware-related cyber-attacks in South African financial institutions. Master's dissertation. Johannesburg: University of the Witwatersrand.

Mugwagwa, T., Phiri, N. & Dube, C. 2024. Cybersecurity issues in South African SMEs: Risks and solutions. Pretoria: University of South Africa Press.

Musa, L. 2023. A systems engineering approach to cybersecurity challenges in South African SMEs. Master's dissertation. Durban: University of KwaZulu-Natal.

Ncubukezi, T. and Mwansa, L., 2021. Best practices used by businesses to maintain good cyber hygiene during Covid19 pandemic. *Journal of Internet Technology and Secured Transactions*, *9*(1), pp.714-721.

Ncubukezi, T., 2022a. Human errors: A cybersecurity concern and the weakest link to small businesses. In Proceedings of the 17th *International Conference on Information Warfare and Security* (p. 395).

Ncubukezi, T., 2022b. Impact of information security threats on small businesses during the COVID-19 pandemic. In *European Conference on Cyber Warfare and Security*, 21(1), pp. 401-410.

Phahlamohlaka, J., Theron, J. & Aschmann, M. 2022. National Cybersecurity Implementation in South Africa: The Conundrum Question. Vol. 21. Pretoria: ArmisteadTec.

Pieterse, H. 2021. The cyber threat landscape in South Africa: A 10-year review. Pretoria: LINK.

Reshmi, T. 2021. Information security breaches due to ransomware attacks – a systematic literature review. Amsterdam: Elsevier Ltd.

Saeed, S., Altamimi, S., Alkayyal, N., Alshehri, E. & Alabbad, D. 2023. Digital transformation and cybersecurity challenges for business resilience: Issues and recommendations. Basel: MDPI.

Shingange, J. 2022. Problematizing the South African cybersecurity policy landscape. Master's dissertation. Stellenbosch: Stellenbosch University.

Tetteh, A.K., 2024. Cybersecurity needs for SMEs. Issues in Information Systems, 25(2), pp.33–42. IACIS.

Tsiodra, M., Panda, S., Chronopoulos, M. & Panaousis, E. 2023. Cyber risk assessment and optimization: A small business case study. Vol. 1. New Jersey: IEEE.

Venkatasubramanian, D., Goyal, S., Ezhilarasan, G., Sharma, Y., Vijayalakshmi, V. & Garg, P. 2024. Evaluating the effectiveness of multi-factor authentication for preventing cyber attacks. Kamand: IEEE.

Biosecure-LLM Framework: Protecting LLMs from Cyberbiosecurity Threats and the Case for Independent AI Safety Governance

Xavier-Lewis Palmer[1], Lucas Potter[1], Srdjan Lesaja[2], Sotirios Karathanasis[3] and Mohammad Ghasemigol[4]

[1]Biosview Labs, Dayton, Ohio, USA
[2]Virginia Commonwealth University, VA, USA
[3]Independent Researcher, New York, NY, USA
[4]Old Dominion University, Norfolk, VA, USA

XP.BV@pm.me
Biosview1@proton.me
srdjanlesaja@gmail.com
sfk@use.startmail.com
mghasemi@odu.edu

Abstract: Large Language Models (LLMs) are becoming critical infrastructure in scientific, healthcare, and governmental contexts. As frontier AI laboratories increasingly partner with government agencies, a fundamental question arises: Who should control the safety and policy-enforcement layers that constrain model behavior? Current safety mechanisms (LLM guardrails) are typically designed for generic "harmlessness" and operate by detecting semantic patterns and refusing requests. However, they are inadequate governance instruments because they cannot implement auditable, domain-specific controls tied to external regulatory policy objects (e.g., control lists or rules governing personally identifying information). Even a perfectly aligned model is not able to express institution-specific policy without an external control layer. This paper argues that the logical separability of policy enforcement from model inference, demonstrated by firewall-style architectures, demands corresponding institutional separability as well. Concentrating both model development and safety governance within the same commercial entities creates unacceptable conflicts of interest, regulatory capture risks, and accountability gaps. We propose that the policy control layers must be housed within independent regulatory bodies, governmental agencies, or trusted third parties rather than the organizations that build and profit from the underlying models. Drawing on the Biosecure-LLM framework as a technical proof-of-concept, we demonstrate that such separation is architecturally feasible and argue it is well-suited for verifiable compliance.

Keywords: AI governance, Institutional design, Regulatory independence, Biosecurity, Responsible AI, Policy enforcement, separability

1. Introduction: The Governance Gap in Frontier AI

Large language models have swiftly moved from research curiosity to embedded infrastructure. In biomedicine, they can assist with literature synthesis, hypothesis generation, and protocol drafting; in clinical settings, they summarize encounters, enable review of clinical histories, and draft documentation for human review (Akogo et al, 2025; Shah, Entwistle, and Pfeffer, 2023; Gao et al, 2024; Floridi et al, 2025). Their integration with electronic knowledge bases and laboratory systems means their behavior now influences scientific and clinical decision-making in ways that impact critical digital infrastructure. Simultaneously, frontier AI laboratories are courting governmental partnerships at an unprecedented pace. These organizations seek contracts that will grant them access to classified information, sensitive health records, and national security data. The commercial incentives are clear: government contracts represent stable revenue and legitimacy. But this interest raises a profound question that existing AI ethics frameworks have overlooked: When the same entity both builds a powerful AI system and controls the mechanisms meant to constrain it, can we trust that there will not be a conflict of interest? This paper argues that the answer is no, not because AI developers are uniquely untrustworthy, but because of institutional design. This concern is compounded when considering the wider bioeconomy: convergence of digital and biological risks, captured by the field of cyberbiosecurity, has been recognized for years (Murch et al., 2018; Potter and Palmer, 2023). LLMs introduce a distinct challenge: because they interface through natural language, they are susceptible to prompt-level manipulation that can subvert safety behaviors. "Prompt injection" and related techniques can induce models to ignore safety instructions, disclose hidden system prompts, or generate content that deviates from institutional policy (Hui et al., 2024). Adversarial prompting and novel techniques in prompt injection can outpace safety considerations and can be difficult to anticipate. Existing alignment techniques, including reinforcement learning from human feedback and instruction tuning, reduce many unsafe outputs but do not necessarily implement enforceable, domain-specific controls mapped to biological regulation or screening (Christiano et

al, 2017; Ouyang et al., 2022; Huang et al, 2025). This creates a policy gap. Traditional cybersecurity focuses on networks and endpoints; traditional biosecurity focuses on physical containment and material screening. The LLM-mediated interface sits between these domains. If left to ad hoc prompt engineering and generic "harmlessness" tuning, institutions risk either over-blocking legitimate discourse or under-blocking content that lowers barriers to harm.

1.1 LLM Guardrail Insufficiency: Current Limitations in Standard LLM Safety Layers

The Biosecure-LLM framework, described below, shows that safety controls can be architecturally separated from model weights. This paper's central claim is that technical separability must be matched by institutional separability: the organizations that enforce AI safety policy must be distinct from those that develop and commercialize the models themselves.

Table 1: The Governance-Safety Mismatch: Why LLM Guardrails are Not Sufficient for Enterprise Compliance: This table maps institutional compliance requirements (governance) to the function of current LLM guardrails (safety). The analysis formalizes the "Structural Gap" as a failure of system integration, entity awareness, and auditability. The inclusion of the "Proposed Architectural Solution" column transitions the argument to a strategic mandate, confirming the necessity of an external, policy-as-code control plane (a Policy Engine) to bridge the gap between general safety alignment and enforceable, auditable enterprise governance (Bai et al, 2022; Mökander et al., 2023; Herrera-Poyatos et al., 2025; Mandalawi et al., 2025; Narula et al., 2025)

Biosecurity/ Compliance Requirement	What Institutions Need (Governance Objective)	What Current LLM Guardrails Do (Safety Objective)	Structural Gap (Why Guardrails Fail Governance)	Proposed Architectural Solution
Regulated-Entity Awareness	Explicit mapping to controlled agents, materials, and techniques (via external, dynamic control lists).	Keyword or latent safety triggers not tied to control lists (detects *intent* to harm, not *policy violation*).	No enforceable policy alignment or entity recognition. Decisions are non-binding.	External Policy Engine with RAG over Control Lists.
Operational Specificity Control	Suppress stepwise, parameterized, optimization content for controlled techniques.	Allow high-level instructions if framed as "educational" or "hypothetical."	Actionability/Plausibility not measured; focuses on high-level *topic* refusal, not *methodology* refusal.	Policy Engine that analyzes response structure (e.g., presence of required parameters, sequence).
PII/PHI/CUI Handling	Dynamic masking, sanitization, or refusal based on data classification and user role (e.g., *only* HR can see salary PII).	Static detection of sensitive *topics* (e.g., "don't share personal medical advice").	LLM lacks integration with existing Role-Based Access Control (RBAC) and Data Loss Prevention (DLP) systems.	Policy Engine integrated via API to RBAC and DLP systems for real-time policy checks.
Auditability	Reproducible, explainable safety decisions (policy rule refusal).	Non-deterministic refusals; black-box model-generated rationales.	No compliance evidence; decisions cannot be traced to a specific, auditable policy rule or external entity.	Policy Engine that logs the specific rule (Policy-as-Code) that triggered the refusal for full compliance trace.

2. The Problem: Concentrated Control and Conflicts of Interest

2.1 The Current Model: Providers as Judge and Jury

Currently, frontier AI laboratories simultaneously develop models, define safety policies, implement enforcement mechanisms, and evaluate their own compliance. This concentration would be remarkable in any other industry. We do not permit pharmaceutical companies to approve their own drugs, nor allow banks to audit their own capital reserves. Yet we accept that AI companies can control every layer of the safety stack. As industries (healthcare, business, transport, etc.) adopt these models, wrapping them in applications for

internal or client-side use, this risk propagates alarmingly fast through the entire critical infrastructure network. The conflict of interest is structural. Safety measures impose costs: they require computational resources, introduce latency, and, most importantly, may restrict model capabilities that drive commercial value. However, a company that controls both the model and its constraints faces constant pressure to relax those constraints when they interfere with user satisfaction or competitive positioning. Even well-intentioned developers have incentives to reward capability expansion and punish excessive caution. This problem intensifies as AI laboratories seek government contracts. Access to sensitive government data creates obligations that conflict with transparency and accountability. When a company trains models on classified information, external oversight becomes nearly impossible due to security constraints. As a consequence, the very partnerships that grant legitimacy to AI developers will shield them from scrutiny.

2.2 Taxonomy of Threats Requiring Independent Oversight

Two families of attacks dominate the risk landscape for LLM deployments: prompt injection and policy evasion. Prompt injection encompasses direct attempts to coerce models into ignoring safety instructions, as well as indirect attacks where adversarial instructions are embedded in retrieved documents that are later fused into the model's context by retrieval-augmented generation pipelines. The latter is particularly insidious because attackers can poison content upstream beyond the LLM's perimeter (De Stefano et al., 2024). Prompt-leaking attacks, another form of prompt injection, aim to extract system prompts and safety rationales; once disclosed, those artifacts enable targeted bypasses (Hui et al., 2024; Asl et al., 2025). Other attacks involve hiding adversarial commands in policy-like format such as JSON or XML, or utilizing fictionalized scenarios to bypass an LLM’s safety content guidelines (Das, 2024; Zhang et al, 2025). Policy evasion involves models producing fluent, operationally specific content (including procedural details, optimization advice, or procurement vectors) that could lower barriers to harmful activity. Even benign queries can elicit fabricated steps that become actionable. This behavior traces to overgeneralization from training data and pressure to provide answers (Farquhar et al., 2024). Findings of AI use being device- and context-dependent by Costa-Gomes et al. (2025) combined with insights of the pervasiveness of AI in wider swathes of community life by Narula et al. (2025) add to an overall picture of a massive attack surface being made available by accidental or malicious actors. These technical vulnerabilities cannot be adequately addressed by the same organizations that profit from model creation and deployment. Detection of novel attack patterns, maintenance of control lists aligned with international regulations, and enforcement decisions that may reduce model utility all require independence from commercial pressures.

2.3 Limitations of Provider-Controlled Safety

Three gaps persist when safety remains provider-controlled. First, a specification gap: generic harmlessness is not an adequate screening threshold for LLMs using complex, critical datasets. A model can be well-aligned in aggregate yet lack enforceable mappings between regulated entities and interventions (Adam et al., 2011; Wheeler et al., 2024). Second, an adversarial gap. Alignment is typically performed on static datasets, while attackers iterate rapidly. Closing this gap requires layered, updateable controls that can be revised as attack patterns evolve, revisions that may not align with provider timelines or commercial interests. Third, an assurance gap. Regulators and institutional risk officers need evidence: tamper-evident logs, provenance records, and reproducible decisions, for example. Conventional alignment pipelines do not provide such artifacts. In adjacent domains, verifiable logging infrastructures demonstrate that append-only, publicly auditable records can raise assurance without exposing sensitive content (Laurie, 2014). An analogous approach for LLM safety requires organizational independence to be credible.

3. Technical Proof-of-Concept: The Biosecure-LLM framework

To demonstrate that institutional separation is technically feasible, we describe Biosecure-LLM, a firewall framework that enforces policy at the boundaries of model communication rather than inside model weights. This architecture proves that safety enforcement can be logically, and therefore institutionally, decoupled from model development.

3.1 Architectural Overview

Biosecure-LLM acts as a security guard for the language model, managing all incoming requests and outgoing answers. When a request comes in from a client application, it first passes through an entrance point, which sends it to the Input Inspection service (Layer 1). This Layer 1 checks the request to make sure it's safe and allowed. If it passes the check, the request is then sent to the language model. The model's answer is then sent to the Output Sanitization service (Layer 2) for cleanup and safety checks before being returned to the client.

This interposition offers three advantages. First, it decouples safety enforcement from model provenance; institutions can upgrade models without rewriting policy. Second, it localizes assurance: the control plane becomes the single locus for audit logging and regulator-facing evidence. Third, it enables comparative assessment: controls act orthogonally to the model, so the same harness can quantify risk reduction across model families (Li et al., 2024; Mazeika et al., 2023; Rein et al., 2023).

3.2 Layer 1: The Prompt Gatekeeper

Layer 1 reduces exposure by vetting user intent and retrieved context before inference through two mechanisms. First, semantic anomaly detection estimates whether an input belongs to a distribution of permitted requests, flagging patterns associated with jailbreaks, instruction overriding, or indirect injection markers (Hendrycks and Gimpel, 2017; Xu et al., 2021). Second, domain-aware recognition identifies mentions of organisms, agents, materials, and techniques, mapping them onto structured policy objects that mirror international control-list categories (Lee et al., 2020; Neumann et al., 2019; Adam et al., 2011; Kobokovich et al., 2019; Wheeler et al., 2024). Entity recognition alone does not trigger enforcement; it conditions a ruleset distinguishing scholarship from attempts to elicit operationally specific procedures. For retrieval-augmented generation, Layer 1 screens not only user strings but retrieved context, rejecting sources that fail provenance requirements, stripping instruction-like segments, and neutralizing patterns known to trigger prompt-leaking (De Stefano et al., 2024; Hui et al., 2024). Another augmentation to Layer 1 involves employing an ensemble model approach, parsing the request through an LLM trained explicitly to recognize and query adversarial requests. Such an adversarial-forward LLM can continue to adapt to emerging prompting attacks while acting as a modular asset that can be paired with different LLMs. Additionally, it can intercept and ask the requester for further information, credentials, and intent, which can help it balance its own weights as it continues to encounter potential attackers.

3.3 Layer 2: The Response Censor

Layer 2 inspects model outputs for their risk of increasing capacity to perform restricted biological activities or are operationally specific (stepwise procedures, parameterized conditions, procurement vectors). This prevents actionable harm, mirroring dual-use governance in nucleic-acid synthesis screening (Adam et al., 2011; Kobokovich et al., 2019; Wheeler et al., 2024). The sanitization pipeline applies three passes: (1) content triage identifying regulated entities; (2) capability inference scanning for procedural markers and optimization advice; and (3) exploitability scoring aggregating operational proximity to controlled entities. If scores exceed certain thresholds, Layer 2 transforms responses by replacing disallowed spans with safe alternatives while preserving legitimate communication (Farquhar et al., 2024). For agentic systems, Layer 2 enforces strict separation between narrative text and executable actions, preventing structured commands from executing without human review (Boiko et al., 2023; Bran et al., 2024).

3.4 Key Technical Properties

The framework emphasizes modularity, low latency, and continuous alignment with evolving policy. Layers are packaged as independent, stateless services that can scale. Policy updates are versioned as digitally-signed, authenticated objects with provenance metadata, enabling rapid rollback and verification (Laurie, 2014; Laurie, 2021). Critically, the framework is model-agnostic. It treats the base LLM as a modular component, regardless of weight visibility or hosting architecture, so long as all inputs and outputs route through the control plane. This preserves diverse deployment choices while maintaining invariant guardrails where they matter which is at the boundary where text could inform action.

4. The Argument for Institutional Separation

4.1 From Logical to Institutional Separability

The Biosecure-LLM framework demonstrates that policy enforcement can be separated from model inference. But separability is not just a technical convenience, but it is an institutional responsibility. Consider the analogy to financial regulation. Markets perform complex transactions that regulators cannot replicate in real-time. Yet we do not conclude that trading firms should therefore regulate themselves. In fact, we have established independent bodies with authority to set standards, conduct audits, and impose penalties. The computational sophistication of the economy does not obviate the need for external oversight. It actually makes such oversight more important. The same logic applies to AI safety. The technical complexity of LLMs is not an argument for provider self-regulation, instead it strongly supports building independent institutions capable of

oversight. The Biosecure-LLM framework shows that policy enforcement can be implemented independent of, and coherently with, model development. Thus, this technical possibility could become an institutional reality.

4.2 Why Model Providers Cannot Own the Policy Layer

There are several discouraging provider-controlled safety regulations. First, accountability requires independence. When incidents occur, investigations must be conducted by parties without a financial stake in the outcome. Provider-controlled safety creates situations where the investigated party itself controls the evidence, sets the standards against which they are measured, and determines what constitutes compliance.

Second, regulatory capture becomes inevitable when regulated entities control enforcement. AI laboratories possess resources, technical expertise, and political access that would dwarf any early regulatory bodies. If providers control the policy layer, they will shape it to accommodate their commercial interests, regardless of stated intentions. Third, unmanaged government partnerships introduce too many conflicts of interest. Frontier laboratories are actively seeking contracts that will grant access to sensitive data. Models trained on classified information, health records, or national security data require oversight that cannot be delegated to commercial entities with competing obligations to shareholders, employees, and customers. Such models must have clear metadata management and control over which users are authenticated to receive responses from models weighted with and without critical or private information. Fourth, public trust requires transparency. Citizens and institutions must be able to verify that safety enforcement serves public interests rather than commercial ones. This verification is less possible when the same organization controls both the model and its constraints.

4.3 Who Should Control the Policy Layer?

Independent oversight could take several forms. Dedicated regulatory agencies with technical capacity and statutory authority represent one model, analogous to the Food and Drug Administration or the Securities Exchange Commission. Alternatively, existing agencies could be augmented with AI safety divisions (an approach adopted in biosecurity) where existing frameworks provide foundations for AI-specific enforcement. Third-party certification bodies offer another model, drawing on precedents in cybersecurity auditing and safety certification. Such bodies could be industry-funded but governance-independent, with statutory backing for their authority. In a similar vein, watchdog-style groups may organically form to publicize breaches in trust.

The above outcomes may emerge at different times or co-exist simultaneously. The form matters less than the principle: the organization that enforces AI safety policy must be structurally independent of organizations that profit from AI deployment. This independence must be visible and verifiable with governance that both prevents capture and enables correction.

5. Governance and Auditability Under Independent Oversight

Independent oversight requires infrastructure to support accountability. The Biosecure-LLM framework couples technical controls with governance mechanisms that demonstrate how such an infrastructure could function.

5.1 Incident Response and Adjudication

When either layer issues an intervention, the system records an immutable event containing a hashed representation of the prompts and context, model and policy versions in force, rules that fired, and outputs that were transformed or suppressed. This snapshot commits to an append-only log so that subsequent policy changes cannot erase the presence and rationale of an executed intervention. Following transparency systems in web public-key infrastructure, inclusion and consistency of records are provable using relatively tamper-proof chains called Merkle-trees: each batch extends the tree, and periodic publication of signed tree heads allows independent verification that past entries have not been altered (Laurie, 2014; Laurie, 2021). The log stores protected elements and structured references rather than plaintext, protecting privacy while enabling audit. Under independent oversight, a designated safety officer receives events for human-in-the-loop adjudication, determining whether interventions were appropriate and diagnosing whether there was attempted misuse, accidental over-specification, or false positives requiring calibration. This separation of duties (automatic detection followed by expert review) mirrors safety controls in critical systems.

5.2 Verifiable Compliance

The logs support third-party verification without revealing sensitive content. Each event stores cryptographic commitments to content digests and policy artifacts. External reviewers can verify that decisions cited specific

policy objects with known versions without accessing underlying text. This approach borrows from secure audit logs and key-transparency systems (Schneier and Kelsey, 1999; Laurie, 2014; Laurie, 2021). Regulators can require periodic submission of signed tree heads and random-audit inclusion proofs as conditions of continued operation, without receiving content triggering confidentiality concerns. Because the control plane externalizes policy, institutions can demonstrate that identical requests would elicit identical interventions across models, a property attractive to compliance officers responsible for multiple portfolios of models.

5.3 Continuous Assessment

Assessment completes the governance loop. The testing framework automatically runs two groups of checks whenever system rules are modified long with routine schedule. The first targets misuse: jailbreak prompts, prompt-leaking patterns, and indirect-injection scenarios exercise suppression of unsafe completions (Hui et al., 2024; De Stefano et al., 2024). Stress suites like Weapons of Mass Destruction Proxy (WMDP) and HarmBench provide baselines for processing at rates that would be questionable by a non-malicious user (Li et al., 2024; Mazeika et al., 2024). The second family targets utility: domain competence on technical questions and performance on benign workloads establish whether scientific capability is preserved and false-positive rates remain acceptable (Rein et al., 2023). Tests run with and without interposition, allowing clean estimates of marginal risk reduction and utility cost. Thus, independent oversight bodies could control assessment standards, conduct periodic assessments, and publish aggregate results.

6. Conclusion and Recommendations

The convergence of powerful AI systems, sensitive biological and governmental data, and nascent regulatory frameworks creates an urgent governance challenge. Frontier AI laboratories are positioning themselves as trusted partners to governments worldwide, seeking access to data that will make their models more capable and their products more valuable. The commercial reasoning is compelling; however, the governance implications are troubling especially as data access grows (Lesaja and Palmer, 2019). This paper has argued that the policy enforcement layer, the mechanisms that constrain AI behavior and ensure compliance with safety standards, must be institutionally separated from model providers. Our goal is not to impede or otherwise handicap AI frontier labs, as their work is valuable. Rather, technical architectures like Biosecure-LLM demonstrate that such separation is feasible and provides benefits for both model creators and data curators. Policy can be enforced at input-output boundaries without retraining models or accessing their weights. The logical separability of enforcement from inference proves that institutional separability is achievable. Concentrated control creates unacceptable risks: conflicts of interest that bias safety decisions toward commercial considerations, accountability gaps that prevent effective investigation of incidents, regulatory capture that subordinates public safety to private profit, and opacity that undermines public trust will erode the efficacy and usability of models. Our recommendations are direct. First, regulatory bodies should assert AI safety policy rather than delegating to providers. The Biosecure-LLM framework provides a template. Second, governmental contracts with AI laboratories should require that safety enforcement be conducted by independent parties, with access to audit logs. Third, industry standards should evolve toward externalized policy enforcement with verifiable compliance mechanisms. The bioeconomy provides instructive precedent. Gene-synthesis screening, control lists for regulated agents, and biosafety review boards all embody the principle that safety oversight must remain independent of commercial interests (Adam et al., 2011; Kobokovich et al., 2019; Wheeler et al., 2024). These frameworks emerged because the alternative (trusting biotechnology companies alone to regulate themselves) poses significant risk of harm. We may expect more regulation as more sensitive biological data, such as radiological and neurophysiological recordings, become easier to store (Lesaja and Palmer, 2020; Liv and Greenbaum, 2023). Work from d'Ascoli et al, (2025) shows that we are increasingly closer to this capability. The same recognition must occur for AI, borrowing from biosafety insights from Gillum et al (2025).

Ethics Declaration: Ethical clearance for the research referred to in this paper was not needed.

AI Declaration: AI tools within Grammarly and Google Docs were used for formatting, grammar/spell-checking, and re-organization.

References

Adam, L., Kozar, M., Letort, G., Mirat, O., Srivastava, A., Stewart, T., Wilson, M.L. and Peccoud, J., 2011. Strengths and limitations of the federal guidance on synthetic DNA. Nature biotechnology, 29(3), pp.208-210. doi: 10.1038/nbt.1802

Akogo, D., Ayensu, J., Sam, N., Hattoh, G., Nyarko, P., Eshun, S., Alhasan, M., Mensah-Brown, H. and Quashie, P., 2025. Moremi Bio Agent: Application of A Foundation Model and End-to-End Automation in the Design and Validation of Monoclonal Antibodies Targeting Plasmodium falciparum Invasion Complex. bioRxiv, pp.2025-02. doi:10.1101/2025.02.12.637967

Asl, J.R., Narula, S., Ghasemigol, M., Blanco, E. and Takabi, D., 2025, November. NEXUS: Network Exploration for eXploiting Unsafe Sequences in Multi-Turn LLM Jailbreaks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 24278-24306). doi:10.18653/v1/2025.emnlp-main.1235

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C. and Chen, C., 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Boiko, D.A. et al. (2023) 'Autonomous chemical research with large language models', Nature, 624, pp. 570–577. doi:10.1038/s41586-023-06792-0.

Bran, A.M., Cox, S., Schilter, O., Baldassari, C., White, A.D. and Schwaller, P., 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, *6*(5), pp.525-535. doi: 10.1038/s42256-024-00832-8

Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. and Amodei, D., 2017. Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

Costa-Gomes, B., Chen, S., Hsueh, C., Morgan, D., Schoenegger, P., Shah, Y., Way, S., Zhu, Y., Adeline, T., Bhaskar, M. and Suleyman, M., 2025. It's About Time: The Temporal and Modal Dynamics of Copilot Usage. arXiv preprint arXiv:2512.11879.

d'Ascoli, S., Bel, C., Rapin, J., Banville, H., Benchetrit, Y., Pallier, C. and King, J.R., 2025. Towards decoding individual words from non-invasive brain recordings. Nature Communications, 16(1), p.10521. doi: 10.1038/s41467-025-65499-0

Das, N., Raff, E. and Gaur, M., 2024. Human-Interpretable Adversarial Prompt Attack on Large Language Models with Situational Context. arXiv preprint arXiv:2407.14644.

De Stefano, G., Schönherr, L. and Pellegrino, G., 2024. Rag and roll: An end-to-end evaluation of indirect prompt manipulations in llm-based application frameworks. *arXiv preprint arXiv:2408.05025*.

Farquhar, S., Kossen, J., Kuhn, L. and Gal, Y., 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, *630*(8017), pp.625-630. doi: 10.1038/s41586-024-07421-0

Floridi, L., Morley, J., Novelli, C. and Watson, D., 2025. What Kind of Reasoning (if any) is an LLM actually doing? On the Stochastic Nature and Abductive Appearance of Large Language Models. arXiv preprint arXiv:2512.10080.

Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J.R., Ektefaie, Y., Kondic, J. and Zitnik, M., 2024. Empowering biomedical discovery with AI agents. Cell, 187(22), pp.6125-6151. doi: 10.1016/j.cell.2024.09.022

Gillum, D.R., Knight, C. and Vogel, K.M., 2025. Understanding biosafety practitioner perspectives. Politics and the Life Sciences, pp.1-24. doi: 10.1017/pls.2025.10014

Hendrycks, D. and Gimpel, K. (2017) 'A baseline for detecting misclassified and out-of-distribution examples in neural networks', arXiv preprint, arXiv:1610.02136. doi:10.48550/arXiv.1610.02136.

Herrera-Poyatos, A., Del Ser, J., de Prado, M.L., Wang, F.Y., Herrera-Viedma, E. and Herrera, F., 2025. Responsible Artificial Intelligence Systems: A Roadmap to Society's Trust through Trustworthy AI, Auditability, Accountability, and Governance. arXiv preprint arXiv:2503.04739.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. and Liu, T., 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), pp.1-55. doi: 10.1145/3703155

Hui, B., Yuan, H., Gong, N., Burlina, P. and Cao, Y. (2024) 'PLeak: Prompt leaking attacks against large language model applications', in Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 390–404. doi:10.1145/3658644.3670370.

Kobokovich, A., West, R., Montague, M., Inglesby, T. and Gronvall, G.K. (2019) 'Strengthening security for gene synthesis: Recommendations for governance', Health Security, 17(6), pp. 419–429. doi:10.1089/hs.2019.0110.

Laurie, B., 2014. Certificate transparency. Communications of the ACM, 57(10), pp.40-46. doi: 10.1145/2659897

Laurie, B., Langley, A., Kasper, E., Messeri, E. and Stradling, R., 2021. Certificate transparency version 2.0. *Internet Requests for Comments, RFC Editor, RFC*, *9162*.

Lee, J., Yoon, W. and Kang, J. (2020) 'BioBERT: a pre-trained biomedical language representation model for biomedical text mining', Bioinformatics, 36(4), pp. 1234–1240. doi:10.1093/bioinformatics/btz682.

Lesaja, S. and Palmer, X.L., 2020. Brain-computer interfaces and the dangers of neurocapitalism. arXiv preprint arXiv:2009.07951.

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J.D., Dombrowski, A.K., Goel, S., Phan, L. and Mukobi, G., 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Liv, N. and Greenbaum, D., 2023. Cyberneurosecurity. In *Policy, identity, and neurotechnology: The neuroethics of brain-computer interfaces* (pp. 233-251). Cham: Springer International Publishing. doi: 10.1007/978-3-031-26801-4_13

Mandalawi, S.A., Mohammed, M.A., Maclean, H., Cakmak, M.C. and Talburt, J.R., 2025. Policy-Aware Generative AI for Safe, Auditable Data Access Governance. arXiv preprint arXiv:2510.23474.

Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B. and Forsyth, D., 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249.

Mökander, J., 2023. Auditing of AI: Legal, ethical and technical approaches. Digital Society, 2(3), p.49. doi: 10.1007/s44206-023-00074-y

Murch, R.S., So, W.K., Buchholz, W.G., Raman, S. and Peccoud, J. (2018) 'Cyberbiosecurity: An emerging new discipline to help safeguard the bioeconomy', Frontiers in Bioengineering and Biotechnology, 6, p. 39. doi:10.3389/fbioe.2018.00039.

Narula, S., Ghasemigol, M., Carnerero-Cano, J., Minnich, A., Lupu, E. and Takabi, D., 2025. Exploring AI Security: A Systematic Mapping Study. IEEE Access. doi: 10.1109/ACCESS.2025.3567195.

Neumann, M., King, D., Beltagy, I. and Ammar, W. (2019) 'ScispaCy: Fast and robust models for biomedical natural language processing', in Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 319–327. doi:10.18653/v1/W19-5034.

Ouyang, L. et al. (2022) 'Training language models to follow instructions with human feedback', arXiv preprint, arXiv:2203.02155. doi:10.48550/arXiv.2203.02155.

Potter, L. and Palmer, X.L., 2023. Mission-aware differences in cyberbiosecurity and biocybersecurity policies: Prevention, detection, and elimination. In Cyberbiosecurity: A new field to deal with emerging threats (pp. 37-69). Cham: Springer International Publishing. doi: 10.1007/978-3-031-26034-6_4

Rein, D., Aurelio, T., Tamkin, A., Ganguli, D. and Steinhardt, J. (2023) 'GPQA: A graduate-level, Google-proof Q&A benchmark', arXiv preprint, arXiv:2311.12022. doi:10.48550/arXiv.2311.12022.

Shah, N.H., Entwistle, D. and Pfeffer, M.A. (2023) 'Creation and adoption of large language models in medicine', JAMA, 330(9), pp. 866–869. doi:10.1001/jama.2023.14217.

Schneier, B. and Kelsey, J., 1999. Secure audit logs to support computer forensics. *ACM Transactions on Information and System Security (TISSEC)*, *2*(2), pp.159-176. doi: 10.1145/317087.317089

Wheeler, N.E. et al. (2024) 'Developing a common global baseline for nucleic acid screening', Applied Biosecurity. doi:10.1089/apb.2023.0034.

Xu, J., Wu, H., Wang, J. and Long, M. (2021) 'Anomaly Transformer: Time-series anomaly detection with association discrepancy', arXiv preprint, arXiv:2110.02642. doi:10.48550/arXiv.2110.02642.

Zhang, S., Zhao, J., Xu, R., Feng, X. and Cui, H., 2025. Output constraints as attack surface: Exploiting structured generation to bypass llm safety mechanisms. arXiv preprint arXiv:2503.24191.

A Comprehensive Cyber Defense Framework for the Indonesian National Armed Forces: Bridging Governance Gaps for National and ASEAN Cyber Resilience

Timothy Shives and Fibriansyah Fatahillah
Naval Post-graduate School, Monterey, California, USA

timothy.shives@nps.edu
fibriansyah.fatahillah.id@nps.edu

Abstract: Indonesia's National Data Center (*PDN*) was targeted by a ransomware attack on June 20, 2024, paralyzing 210 government agencies, causing manual immigration procedures, and exposing significant weaknesses in Indonesia's cyber governance system. The National Cyber and Crypto Agency (*BSSN*) was mandated under Presidential Regulation 47/2023 to coordinate the response, but the response operation remained disorganized due to various agencies working independently without a unified leadership system, including the Indonesian National Armed Forces (*TNI*) operating independently despite possessing a Cyber Unit (*Satsiber*) with adequate cyber warfare capabilities. The attack on the *PDN* ultimately revealed three governance weaknesses: a lack of a unified command system for conducting national-scale response operations, the separation of military resources from the protection of civilian infrastructure, and a systemic failure to maintain adequate operational readiness. Through a comparative analysis of cyber command models in the United States, Singapore, South Korea, and Australia, combined with an institutional assessment using the McKinsey 7S and NIST frameworks, we propose an integrated defense architecture. The establishment of a Joint Cyber Defense Task Force (JCDTF) operating under a proposed civilian-military organization, the National Cyber Security Coordination Center (NCCC), would create a single command system for crisis response and maintain democratic civilian control through established legal authority, mandatory parliamentary oversight, and limitations on operational areas. This framework would address existing governance weaknesses through democratic cyber governance principles that can also be used by ASEAN countries to address their civil-military integration challenges in handling national-scale cyber incidents.

Keywords: Cyber resilience, Cyber governance, Civil-military relations, Indonesia, ASEAN, Crisis management

1. Introduction

A cyber incident was discovered when immigration officers at Soekarno-Hatta International Airport found their systems encrypted by ransomware upon arriving at the airport at 8:00 AM on June 20, 2024 (Azhar, 2024). The attack, carried out by Brain Cipher, caused 210 government agencies to lose access to critical database systems as it spread through Indonesia's National Data Center (*PDN*) within hours (Antoniuk, 2024). The government refused to pay the $8 million ransom demand but required several weeks to recover their data by performing a manual cloud migration to a commercial cloud service (Karmini, 2024; Nugroho, 2024). The attack revealed that governance weaknesses extended beyond the actual technical intrusion, as various agencies failed to coordinate their response to the cyberattack, there was no single command center, and Indonesia's main defense organization, the Indonesian National Armed Forces (*TNI*), was not included in the national cyber emergency response system (Nugroho, 2024).

This research aims to determine what kind of institutional framework can integrate military cyber operations into Indonesia's defense system while maintaining civilian control over the government. The *PDN* attack provides evidence to support this research by demonstrating specific governance weaknesses that necessitate changes in the current organizational structure. The Indonesian government has introduced a draft Cyber Security and Resilience Bill (*RUU KKS*) intended to create well-defined institutions capable of addressing future cyber challenges, but the bill faces challenges of institutional fragmentation due to the lack of a unified operational body (Lebang, 2025).

We propose a Joint Cyber Defense Task Force (JCDTF) operating within the National Cyber Security Coordination Center (NCCC) as an integrated architecture to address the identified failures. This framework aims to maintain operational efficiency while ensuring democratic oversight through defined legal authority, a civilian-military management system, and programs to develop the technical skills of personnel in the cyber domain.

2. Fragmented Indonesian Cyber Governance

Indonesia's current cybersecurity institutions function as separate entities that appear to compete rather than cooperate. The National Cyber and Crypto Agency (*BSSN*) began operations in 2017 based on Presidential Regulation 47/2023, which established an organization to coordinate activities related to cryptography and cybersecurity, but limited its operational scope to only 400 staff working on policymaking (SSEK Law Firm, 2024).

Meanwhile, the Ministry of Communication and Information (*Kemkomdigi*), as the body operates the government's IT infrastructure through the National Data Center (*PDN*), operates its own security management system without cooperation with *BSSN* (Lebang, 2025). On the other hand, the Indonesian National Police (*Polri*) use its Cyber Crime Unit to conduct investigations into cybercrimes. And the State Intelligence Agency (*BIN*) conducts cyber intelligence operations through its own protected and confidential operational system (IISS, 2021).

TNI maintain their own cyber operations through the *TNI* Cyber Unit (*Satsiber*), established in 2017 to defend military C4ISR networks with approximately 150-200 personnels (IISS, 2021; Priyandita & Lebang, 2025). The military follows a separate command structure because Indonesia placed civilian control over its armed forces after the democratic transition in 1998. This separation between civilian and military organizations supports democratic development, but it creates operational weaknesses when cyberattacks affect both sectors (Putra et al., 2022).

Reflecting on the *PDN* incident, Indonesia is proposing a draft Cybersecurity Bill (*RUU KKS*)in 2025 that would give *BSSN* new powers to control content and manage AI systems as well as conduct criminal investigations, but this would create jurisdictional disputes with *Kemkomdigi* and *Polri*, while the *TNI*'s responsibility to participate in national cyber defense remains unclear (Lebang, 2025). Legislative methods address regulatory issues, but do not solve the problem of how to integrate both civilian and military operations for national-scale cyber crisis response management.

3. Theoretical Framework

Crisis management theory defines a crisis as an urgent threat that endangers the core elements of a system, while leaders must make decisions under conditions of uncertainty and time constraints (Boin et al., 2016). Effective response requires organizations to perform five critical tasks: situational awareness to understand the circumstances, decision-making to choose actions, coordination for alignment among agencies, information dissemination to the public, and learning for the implementation of lessons learned. The *PDN* attack met all the requirements for a crisis, but Indonesia's disorganized governance system made the execution of these tasks difficult.

A governance system composed of many complex elements requires coordination that goes beyond the process of distributing official positions (Peters, 2015). The lead agency model achieves success when organizations understand each other's needs and the coordinator possesses the appropriate knowledge, and there are established relationships between the relevant parties, and all teams have experience with established procedures. Presidential Regulation 47/2023 designated *BSSN* to coordinate cyber-related activities, but the regulation was not accompanied by efforts and processes to create the fundamental elements for *BSSN* to function as a lead agency, leading to a situation where coordination exists only in name, but operations remain separate.

Civil-military integration in the cyber domain faces unique challenges related to translation issues. Boeke (2015) points out that kinetic warfare doctrines fail to translate effectively into cyber operations because these operations operate without clear territory and are difficult to trace to their origins, and do not distinguish between military and civilian targets (Dunn Cavelty, 2014; Rid, 2012). Democratic control systems require civilians to acquire technical capabilities that allow them to conduct successful oversight, rather than allowing the armed forces to operate independently due to information gaps (Feaver, 2005). The Indonesian government faces a dilemma between using its armed forces for national cybersecurity defense and protecting the democratic constitution from erosion that could occur through military expansion into internal surveillance activities.

4. Methodology

This research combines several qualitative research methods, including case study analysis, comparative research, and institutional framework development. The *PDN* attack revealed critical evidence of governance weaknesses requiring fundamental organizational change. The attack timeline and inter-agency response patterns emerged from process tracing that analyzed documents from open sources including government press releases, Indonesian and international media coverage, technical security analyses, and academic commentary. Researchers used various sources to confirm the precise timing of events, where they coded data for timestamps, actor attribution, type of action, and outcomes.

This research evaluates the integration of cyber command through a comparative study of four democratic countries that demonstrate advanced integration models and have strategic importance for Indonesia: the United States, Singapore, South Korea, and Australia. The analysis revealed common operational methods for civilian supremacy but showed two distinct command structures and specific elements that influence the ability to transfer knowledge.

The framework synthesis method applied the McKinsey 7S organizational assessment to identify weaknesses in the Indonesian National Armed Forces' cyber capabilities by analyzing the dimensions of strategy, structure, systems, shared values, skills, staff, and style (Waterman et al., 1980). And the NIST Cybersecurity Framework 2.0 established the operational architecture supporting the proposed institutional design (National Institute of Standards and Technology, 2024). This analysis relies on publicly available data due to a lack of access to classified assessments, internal communications, and complete forensic evidence. The proposed framework requires empirical verification through stakeholder feedback collection, scenario testing, and legal assessment of its compatibility, which should be studied by future researchers.

5. *PDN* Attack: A Governance Crisis

5.1 Timeline and Impact of the Attack

The Brain Cipher ransomware variant, linked to the LockBit family, encrypted the *PDN* system on June 20, 2024, affecting 210 government institutions (Antoniuk, 2024). The sudden shutdown of immigration services forced staff to process passports manually, leading to increasingly long delays at international airports (Reuters, 2024). The government decided to refuse the ransom payment on June 24 after four days of negotiations, but the recovery process faced various challenges that lasted for more than four weeks (Karmini, 2024; Meyer, 2024). Government organizations undertook an emergency cloud migration with a commercial provider because their backup systems were inadequate, resulting in data security threats and exposing their weak crisis recovery capabilities (Nugroho, 2024). Technical analysts indicated that the attackers must have maintained a connection for several days before they activated the encryption, as they needed time to collect data and distribute their ransomware throughout the system infrastructure, proving that the government's monitoring systems failed to detect the breach (Antoniuk, 2024).

5.2 Critical Governance Failures

The first critical governance failure occurred due to the lack of an organized command structure during the incident and system recovery. Although *BSSN* had been designated as the national cyber coordinator, publicly available information indicates that the response system remained uncoordinated and siloed. During the crisis, *BSSN's* role remained largely invisible to the public, in contrast to officials from the Ministry of Communication and Information Technology (*Kemkomdigi*) who received most of the media attention for their work in recovery initiatives (Azhar, 2024). The government took four days to announce its refusal to pay the ransom, indicating that they lacked established policies and protocols for such situations. There is no evidence to suggest that *BSSN* directed inter-agency operations or formed a unified crisis command. The agencies released separate public statements instead of issuing a single message through a coordinated effort, further proving the failure of their coordination system despite having an official framework for collaboration on national cyber issues. *BSSN*, in its position as the official control authority, was deemed lacking in the technical capabilities to direct the handling of the cyber incident, leading several institutions to operate independently because they possessed sufficient resources to manage infrastructure systems (*Kemkodigi*) and conduct investigations (*Polri*).

The second failure was the exclusion of the Indonesian National Armed Forces (*TNI*) from the national response. *TNI* possesses cyber capabilities capable of providing network defense capabilities and forensic analysis and incident response expertise to defend their military C4ISR systems (IISS, 2021). Public records contain no evidence that *TNI* personnel were involved in the *PDN* incident response operations, primarily because the institution operates independently from the democratic changes that occurred after 1998, not due to a lack of capability. Indonesia lacks established legal procedures to allow the military to assist civilian infrastructure during a national cyber emergency. The military's response to and involvement in natural disasters through search and rescue assistance demonstrates an anomaly in the cyber domain, as defense institutions are not involved in protecting national infrastructure. Military technical expertise becomes useless as civilian agencies struggle to recover from such events, illustrating how separate organizational structures lacking operational integration result in additional costs and national-scale effects.

The government failed to achieve operational readiness, leading to a third major failure, where the attacks revealed that the government failed to build adequate protection against its various operational vulnerabilities.

Monitoring systems were unable to identify the attackers, who remained in the system for several days during the reconnaissance phase until end-users reported the problem, rather than the central cybersecurity operations center detecting it. Backup systems failed to function properly, requiring organizations to spend weeks recovering their data instead of using backups for rapid recovery. The ad-hoc approach to commercial cloud migration demonstrated a lack of a disaster recovery plan, and there is no public evidence that previous joint exercises to train inter-agency coordination during cyber crises had a positive effect in accelerating reaction or recovery time from incidents. The recovery period, lasting several weeks, indicates that Indonesia has focused its efforts more on developing regulatory frameworks than on building operational resilience capabilities.

Table 1: International Cyber Command Integration Models

Dimension	United States	Singapore	South Korea	Australia	Indonesia Implication
Command Structure	Unified USCYBERCOM dual-hatted with NSA	Separate military branch (DIS) coordinated with CSA	Military subordinate to NIS oversight	Federated across services with DSPF principles	Unified command reduces fragmentation seen in *PDN* response
Civilian Oversight	Congressional committees plus DoD civilian leadership	Cabinet-level coordination through Total Defence framework	Parliamentary reporting requirements to NIS committees	Defence Minister authority through DSPF	Multiple oversight mechanisms prevent military autonomy
Legal Authority	Title 10 (military) versus Title 50 (intelligence) separation	Defence Act amendments creating DIS statutory basis	National Intelligence Service Act provisions	Defence Act plus DSPF framework	RUU KKS requires explicit *TNI* cyber role authorization
Operational Doctrine	Persistent Engagement and Defend Forward proactive posture	Total Defence integration across domains	Offensive capability development for deterrence	Principles-based mission-adaptive approach	Doctrine development must precede structure creation
Personnel Management	Cyber Mission Force with specialized career tracks	DIS direct recruitment with hybrid civil-military staffing	Military cyber specialist retention programs	Flexible civilian-military teaming arrangements	*TNI* needs dedicated cyber career path preventing rotation losses

6. International Models for Integrated Cyber Command

The comparative analysis shows different methods which allow military cyber operations to function under civilian authority while maintaining democratic oversight. The United States Cyber Command achieves unified command effectiveness through its single authority which controls offensive and defensive operations and intelligence functions while operating under civilian Department of Defense leadership and congressional oversight (US Cyber Command, n.d.). The Digital and Intelligence Service of Singapore operates as a regional military cyber force model which unites with the civilian Cyber Security Agency through Total Defence framework to achieve national defense coordination (Fitriani, 2023; Singapore MINDEF, 2025). The National Intelligence Service of South Korea controls military cyber operations through its oversight system which requires parliamentary disclosure of all activities (Wood, 2024). The Australian Department of Defence (2025) established the Defence Security Principles Framework which uses principles to guide decision-making instead of creating specific rules that allow military operations to adjust while staying within limits set by civilian authorities.

The research reveals multiple recurring patterns which appear throughout all studied cases (Table 1). The four democracies uphold civilian control through separate institutional systems which include US legislative oversight and Singapore and Australia use ministerial control and South Korea depends on intelligence agency coordination (Feaver, 2005). The United States and Singapore operate under unified command systems which differ from Australia's federated system because Indonesia's *TNI* joint command structure appears to support a centralized command structure (Indonesian Ministry of Defence, 2015). The development of operational doctrine before or at the same time as organizational structures proved essential for successful cases which makes Indonesia's current lack of doctrine its most critical problem (Priyandita & Lebang, 2025)

7. Proposed Framework: JCDTF and NCCC

7.1 Diagnostic Assessment of *TNI* Cyber Capabilities

The McKinsey 7S analysis shows that *TNI* cyber mission operations do not align with the current organizational structure because of seven different factors (Table 2). The assessment shows that structural changes by themselves do not solve problems because organizations need to transform their entire strategy and culture and workforce structure.

Table 2: McKinsey 7S Assessment of *TNI* Cyber Readiness

Element	Current Status	Critical Gap
Strategy	Reactive perimeter defense	Lacks Defend Forward or active defense doctrine; focus solely on C4ISR protection (Priyandita & Lebang, 2025)
Structure	Siloed from national governance	Cyber Unit isolated from *BSSN* decision-making; limited integration with civilian intelligence
Systems	Fragmented coordination	No unified incident response platform with civilian CERTs; manual coordination during *PDN* attack
Shared Values	Conventional warfare culture	Institutional prioritization of kinetic operations over digital domain; cyber viewed as support function
Skills	Critical shortages	Deficit in forensic analysts and reverse engineers; reliance on generalist IT staff versus specialists
Staff	Understrength at ~40% manning	Rigid rotation policies prevent specialization retention (Antara News, 2025)
Style	Hierarchical command	Top-down structure slows decision-making speed required for cyber incident response

7.2 National Cybersecurity Coordination Centre (NCCC)

The NCCC would serve as a single operational command center which would handle national cyber emergencies to solve the current problem of insufficient coordination that occurred during the *PDN* attack (Nugroho, 2024). The senior civilian director at *BSSN* or ministerial level would take charge of the operation while deputy directors from *TNI* (JCDTF Commander), *BIN*, *Polri* and *Kemkomdigi* would provide support. The *RUU KKS* amendments through statutory authority would allow crisis management during declared cyber emergencies by controlling member agencies while preserving their regular duties for planning and exercise activities and information exchange (Lebang, 2025).

The NCCC must deliver reports to the National Security Council which should contain quarterly briefings for parliament members together with classified operation information in distinct sections (Wood, 2024). Organizations can protect their operational security through annual public threat assessments which also enable them to share complete information with the public (Australian Department of Defence, 2025). The established framework solves *BSSN*'s present authority-capability problem through its creation of legal command authority which draws operational strength from member agency capabilities (SSEK Law Firm, 2024).

7.3 Joint Cyber Defense Task Force (JCDTF)

The JCDTF operates as *TNI*'s primary cyber force which operates through two different command systems: The *TNI* Commander exercises tactical control for military operations, and NCCC maintains operational control for supporting civilian infrastructure (Boeke et al., 2015). The dual subordination system enables organizations to merge their operational capabilities with democratic supervision (Feaver, 2005).

The organizational structure includes three specific elements which follow the US Cyber Mission Force architecture design but maintain proportions suitable for this organization (US Cyber Command, n.d.). The National Defense Teams (NDT) would create fast response systems to assist civilian organizations during major emergencies including the *PDN* attack through NCCC-approved activation for forensic analysis, data recovery and critical infrastructure hardening (Nugroho, 2024). The Combat Support Teams (CST) would unite cyber warfare capabilities with standard *TNI* military operations which would span all three branches of the military including Army and Navy and Air Force (Indonesian Ministry of Defence, 2015). The Force Protection Teams (FPT) would protect *TNI*'s fundamental mission to defend military C4ISR networks through their work of network monitoring and vulnerability testing and threat intelligence gathering and security operation management (IISS, 2021).

The organization will consist of 60% military personnel and 30% civilian technical experts and 10% contractors. The hybrid model solves talent shortages through its ability to create local expertise (Abdurrachman et al., 2024). Personnel management reforms prove essential because they will establish dedicated cyber career paths which extend service periods to 4-5 years instead of typical 2-year assignments to maintain specialized skills and offer 1.5 times standard pay rates to draw candidates and allow civilian experts to join through lateral entry programs and universities can establish recruitment pipelines.

The explicit authorities together with defined limitations serve to stop the expansion of mission scope (Dunn Cavelty, 2014). JCDTF would have complete power to protect *TNI* networks while it would only support civilian infrastructure through *BSSN* Director authorization after receiving a request from NCCC. The organization would perform offensive cyber operations only through presidential authorization which requires National Security Council evaluation (Lebang, 2025). The organization would work with *BIN* to gather intelligence, but *BIN* would keep its main authority to collect intelligence (IISS, 2021). The organization would help *Polri* with technical support during their investigations, but law enforcement agencies would continue to handle all prosecution matters. The established boundaries enforce civilian primacy through direct legal restrictions which replace the need for voluntary self-control (Feaver, 2005).

7.4 Implementation Requirements

The legal system requires RUU KKS to undergo amendments which will establish NCCC as a crisis command center and JCDTF for military network defense and civilian support operations and to create parliamentary oversight systems and prohibit all unauthorized offensive operations (Lebang, 2025). The total budget for five years of implementation needs to be prepared which includes establishment costs for personnel recruitment, training, secure infrastructure, systems, and tools also for annual steady-state operations for personnel, operations, and for modernization. The funding sources for this initiative consist of *TNI* budget reallocation and newly appropriations, international partnerships with US, Australia, South Korea or Singapore and commercial threat intelligence cooperation (Fitriani, 2023).

The NCCC needs to conduct quarterly joint exercises to check its operational readiness because these exercises help assess its coordination systems and create crisis management protocols which determine decision roles and communication methods and set up vital backup systems for infrastructure and continuous improvement through learned lessons. The preparedness investments work to solve the readiness problems which *PDN* attack revealed (Nugroho, 2024).

8. Discussion

8.1 Addressing Democratic Governance Concerns

The military needs to operate within its authorized boundaries because Democratic control mechanisms have created various protective systems. Civilian NCCC director maintains operational authority over JCDTF civilian support missions. The parliamentary oversight committee conducts thorough reviews through its internal technical staff instead of performing typical inspection procedures. The law establishes two types of restrictions which block both home-based surveillance activities and self-operating military attacks. The organization maintains transparency through its annual public reports which show its operational activities and budget usage and capability growth. The protection system for whistleblowers enables staff members to report violations while guaranteeing they will not experience any form of retaliation. The United States and Australia along with other democratic cyber forces operate as international partners which demonstrate their professional military culture through their commitment to civilian control. The system requires structural controls, but these controls do not fully meet all requirements. *TNI* needs professional military education which teaches soldiers about democratic governance through civil-military relations and leadership development that focuses on self-control and responsibility instead of independence. The military should use international exercises to teach its members about democratic values through socialization programs. The process of building civilian capacity stands as an equally important requirement because *BSSN* technical workforce development through training and recruitment and retention practices enables effective oversight instead of information-based rubber-stamping (Aulianisa & Indirwan, 2020). The assessment process receives independent evaluation from technical advisory boards which consist of academic and private sector experts.

8.2 Limitations and Future Research

The framework solves governance deficiencies which researchers discovered through their analysis of *PDN* attack data and their comparison of different systems but needs testing in real-world settings. The research

agenda for future development needs to conduct semi-structured interviews with *TNI*, *BSSN*, BIN and Polri officials which will help assess framework usability and identify barriers to execution. The operational readiness assessment needed by conduct tabletop exercises which will stress test the NCCC coordination on crisis to verify system performance before deploying the full system. The legal assessment of constitutional compatibility with public law evaluation of dual-reporting military structures would protect democratic legitimacy. The analysis of political economy would help identify which organizations oppose authority consolidation and what motivates them to change their opposition which would guide the implementation process.

The framework also needs political support to achieve long-term success because its success requires continuous implementation after the initial establishment of its structural components. The democratic institutions of Indonesia need to maintain their focus on civilian oversight because military autonomy becomes more attractive because of technical difficulties. The analysis of regional effects needs to evaluate ASEAN norms which include non-interference so Indonesia should maintain transparency through its participation in the ASEAN Defence Ministers' Meeting (ADMM) cyber working group, its conduct of defensive military exercises and voluntary assessment programs instead of forcing integration models on neighboring countries with different political structures.

9. Conclusion

The June 2024 *PDN* ransomware attack revealed three core weaknesses in Indonesia's cyber defense governance system because the country operated without a single command center despite creating coordination protocols and because it did not combine military assets with civilian infrastructure protection and because its operational readiness remained insufficient. The institutions failed to stop these disasters because they dedicated their efforts to managing regulations instead of creating proper crisis management infrastructure. The US, Singapore, South Korea and Australia have different models which could prove that within democracies system they can merge their military cyber forces into their existing structures while keeping civilian control. The JCDTF/NCCC framework which the authors propose implements these principles into Indonesia's current institutional framework by having NCCC lead operations under civilian control while JCDTF reports to both organizations. The operationalization of democratic accountability depends on four elements which include explicit legal authorities and parliamentary oversight and prohibited capabilities and civilian technical capacity building.

The research evaluates *PDN* governance failures through complete assessment while it evaluates different integration frameworks to develop specific organizational designs which fix the identified weaknesses. The framework requires empirical testing through stakeholder participation and scenario-based evaluation to confirm its design methodology which derives from theoretical concepts. The success of implementation depends on continuous political support which must focus on developing capabilities and maintaining democratic monitoring systems.

The institutional framework of Indonesia needs to establish a system which unites operational performance with public oversight for digital sovereignty protection. The *PDN* attack created an opportunity for Indonesia to use for protecting its national cyber security systems while showing democratic leadership in ASEAN's digital transformation process. The dialogue needs to move away from emergency response work toward building permanent institutions.

Acknowledgements

The authors express sincere gratitude to the Naval Postgraduate School for academic guidance and resources, and to LPDP for sponsorship.

Ethics Declaration: This research was conducted using publicly available data, including government reports, academic papers, and news articles. As such, it did not involve human subjects or confidential data, and formal ethical clearance was not required.

AI Declaration: Open-source artificial intelligence tools were used solely for proofreading and citation formatting. All analysis, arguments, and conclusions presented in this paper are the original work of the authors or are derived from appropriately cited sources.

References

Abdurrachman, F., Suharjo, B., & Biantoro, Y. (2024). Building The TNI's Defense Posture In Cyberspace: Strategies For Dealing With Cyber Warfare In The Digital Era. IJPSAT, 48(1), 451–461.

Antara News. (2025, February 3). TNI to continue cyber defense training for soldiers. https://en.antaranews.com/news/343530/tni-to-continue-cyber-defense-training-for-soldiers
Antoniuk, D. (2024, June 24). Indonesia's national data center encrypted with LockBit ransomware variant. The Record. https://therecord.media/indonesia-national-data-centre-hacked
Aulianisa, S. S., & Indirwan, I. (2020). Critical Review of the Urgency of Strengthening the Implementation of Cyber Security and Resilience in Indonesia. Lex Scientia Law Review, 4(1), 33–48. https://doi.org/10.15294/lesrev.v4i1.38197
Australian Department of Defence. (2025, May 30). Defence Security Principles Framework | Defence. https://www.defence.gov.au/business-industry/industry-governance/defence-security-principles-framework
Azhar, M. (2024, June 25). Cyberattack on Indonesia's national data centre paralyses government services [Online post]. GovInsider. https://govinsider.asia/intl-en/article/cyberattack-on-indonesias-national-data-centre-paralyses-government-services
Boeke, S., Heinl, C. H., & Veenendaal, M. A. (2015). Civil-military relations and international military cooperation in cyber security: Common challenges & state practices across Asia and Europe. 2015 7th International Conference on Cyber Conflict: Architectures in Cyberspace, 69–80. https://doi.org/10.1109/CYCON.2015.7158469
Boin, A., 'T Hart, P., Stern, E., & Sundelius, B. (2016). The Politics of Crisis Management: Public Leadership under Pressure (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781316339756
Dunn Cavelty, M. (2014). Breaking the Cyber-Security Dilemma: Aligning Security Needs and Removing Vulnerabilities. Science and Engineering Ethics, 20(3), 701–715. https://doi.org/10.1007/s11948-014-9551-y
Feaver, P. (2005). Armed servants: Agency, oversight, and civil-military relations (1. paperpack ed). Harvard Univ. Press.
Fitriani. (2023, September 5). New structures for ASEAN cyber defence and information sharing [Online post]. IISS. https://www.iiss.org/online-analysis/online-analysis/2023/09/new-structures-for-asean-cyber-defence-and-information-sharing/
IISS. (2021). Cyber Capabilities and National Power—Indonesia (pp. 143–151). International Institute for Strategic Studies.
Indonesian Ministry of Defence. (2015). Indonesian Defence White Paper 2015 (Third). Defence Ministry of the Republic of Indonesia. https://www.kemhan.go.id/wp-content/uploads/2016/05/2015-INDONESIA-DEFENCE-WHITE-PAPER-ENGLISH-VERSION.pdf
Karmini, N. (2024, June 24). Indonesia won't pay an $8 million ransom after a cyberattack compromised its national data center. AP News. https://apnews.com/article/indonesia-ransomware-attack-national-data-center-213c14c6cc69d7b66815e58478f64cee
Lebang, C. G. (2025, March 18). Indonesia's Cyber Security and Resilience Bill: Strengthening Governance or Expanding Institutional Rivalries? Https://Www.Lab45.Id/. https://www.lab45.id/detail/298/indonesia-rsquo-s-cyber-security-and-resilience-bill-strengthening-governance-or-expanding-institutional-rivalries
Meyer, C. (2024, June 26). Indonesia Refuses to Pay $8M Ransom in Data Center Cyberattack [Online post]. ASIS International. https://www.asisonline.org/security-management-magazine/latest-news/today-in-security/2024/june/indonesia-ransomware/
National Institute of Standards and Technology. (2024). The NIST Cybersecurity Framework (CSF) 2.0 (No. NIST CSWP 29; p. NIST CSWP 29). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.29
Nugroho, Y. (2024, August 8). Indonesia's National Data Centre Ransomware Attack: A Digital Governance Failure? FULCRUM. https://fulcrum.sg/indonesias-national-data-centre-ransomware-attack-a-digital-governance-failure/
Peters, B. G. (2015). Pursuing Horizontal Management: The Politics of Public Sector Coordination. University Press of Kansas. https://doi.org/10.1353/book46028
Priyandita, G., & Lebang, C. G. (2025, April 10). Indonesia's cyber soldiers: Armed without a compass. The Strategist. https://www.aspistrategist.org.au/indonesias-cyber-soldiers-armed-without-a-compass/
Putra, T. M., Midhio, I. W., & D.A.R, D. (2022). The Challenges of Civil-Military Cooperation in The Face of Cyber Threats in Indonesia. International Journal of Research and Innovation in Social Science, 06(02), 471–474. https://doi.org/10.47772/IJRISS.2022.6221
Reuters. (2024, June 24). Cyber attack compromised Indonesia data centre, ransom sought. https://www.reuters.com/technology/cybersecurity/cyber-attack-compromised-indonesia-data-centre-ransom-sought-reports-antara-2024-06-24/
Rid, T. (2012). Cyber War Will Not Take Place. Journal of Strategic Studies, 35(1), 5–32. https://doi.org/10.1080/01402390.2011.608939
Singapore MINDEF. (2025, July 27). Cyber Defence. Ministry of Defence. https://www.mindef.gov.sg/defence-matters/defence-topics/cyber-defence/
SSEK Law Firm. (2024, May 20). Fortifying Indonesia's Cyber Defenses: New Regulations for National Security and Crisis Management. SSEK Law Firm. https://ssek.com/blog/fortifying-indonesias-cyber-defenses-new-regulations-for-national-security-and-crisis-management/
US Cyber Command. (n.d.). US Cyber Command History. Retrieved December 8, 2025, from https://www.cybercom.mil/About/History/
Waterman, R. H., Peters, T. J., & Phillips, J. R. (1980). Structure is not organization. Business Horizons, 23(3), 14–26. https://doi.org/10.1016/0007-6813(80)90027-0
Wood, N. (2024, August 2). South Korea's 2024 Cyber Strategy: A Primer | Strategic Technologies Blog | CSIS [Online post]. CSIS. https://www.csis.org/blogs/strategic-technologies-blog/south-koreas-2024-cyber-strategy-primer