Towards Trustworthy AI-Based Military Cyber Operations

Clara Maathuis
Open University of the Netherlands, Heerlen, The Netherlands

clara.maathuis@ou.nl

Abstract: Within the dynamic realm of contemporary warfare, Artificial Intelligence (AI) emerges as a transformative force that reshapes the ways and means used to strategize, execute, and assess military operations. In this journey, the use of AI spans functions and capabilities like intelligence analysis, target engagement decision-making support, weapon autonomy, and effects analytics. Concurrently, AI enhances, e.g., the effectiveness of military plans and capabilities having the potential to reducing risks to civilians, civilian objects, and military personnel. In this rapidly evolving arena, military Cyber Operations gained unprecedented prominence due to their intrinsic digital and cross-domain nature, speed, and became a clear option to achieving military goals, and a mature set of alternatives to conventional ones. Nonetheless, they need continuous assessment, deal with different uncertainty types produced by characteristics like anonymity and can imply psychological impact. Hence, such military operations demand meticulous planning, sophisticated execution, and a deep understanding of technical, military-legal, ethical, and strategic implications and consequences. This represents a direct call for building solutions that align the potential of AI with the responsible and safe conduct of military operations in the military cyber domain: building trustworthy AI-based military Cyber Operations. While incipient efforts to tackle important dimensions of such an approach exist in this domain, a direct and unified approach that unifies them as a commitment and artefact lacks. To tackle this knowledge gap, this research aims to build a bridge between the above-mentioned dimensions by proposing a working definition and framework for building trustworthy AI-based military Cyber Operations using the Design Science Research methodology.

Keywords: Trustworthy AI, Responsible AI, Cyber operations, Cyber security, Military operations

1. Introduction

"Trust but verify." (Ronald Reagan)

Artificial Intelligence (AI) is widely regarded as a complex revolutionary technology fuelled by powerful combinations of data, knowledge, computational strategies, techniques, and resources. Referring to building data, knowledge, or a combination between data and knowledge intelligent systems (i.e., hybrid AI), AI increased in being used in various domains due to its high-performance results, representation power, and decision-making support transforming both society and people's lives. In the military operations realm, AI emerges as a significant force by offering unprecedent capabilities in various activities and actions in tasks (Szabadföldi, 2021; UK DoD, 2022) like enhancing situation awareness in drone reconnaissance missions, enabling realistic decision-making support for target engagement using a malware-based cyber weapon, and providing dynamic/adaptive risks assessments for collateral effects (Maathuis, Pieters & van den Berg, 2018b).

In the military cyber domain, AI-based systems already show their potential in activities like target identification, developing proactive capabilities with built-in impact assessment, and optimizing incident response strategies to adversaries' action(s). Nevertheless, due to its unknown technical, social, and ethical risks and implications, the overall public discourse stimulates fear in society and relevant decision makers in this domain. The development and use of AI systems should be done in contexts and situations characterized by clear, well-structured, and adaptive programs/policies firmly respecting democratic rights and well-being (EU Commission, 2019; Harrison & Luna-Reyes, 2022). This implies translating these concerns in a realistic way considering relevant legal, social, and ethical efforts captured through norms, principles, and values that need to be thought, implemented, and evaluated during all life cycle development phases of AI systems used. The core challenge here lies in understanding and building trustworthy AI systems in military Cyber Operations that are safe, responsible, and robust (EU Commission, 2019; Morgan et al., 2020) by mitigating possible risks and harm produced. Accordingly, a socio-technical approach is necessary for building corresponding technical methods and metrics for each of the aspects pointed in the corresponding legal, social, and ethical norms, principles, and values. This corresponds to countering both the complexities and dynamics of a specific context while accounting the goals and expectations of the stakeholders involved when building/using AI-based military Cyber Operations. For instance, Explainable AI for assuring or enhancing transparency, Responsible AI for supporting accountability, security and privacy assessment mechanisms, verification methods for assuring system robustness (Morgan et al., 2020; Maathuis & Chockalingam, 2023). This calls for a paradigm shift in this domain and to the best of our knowledge, in the military Cyber Operations area, this represents a knowledge gap. Tackling this gap would minimize misconceptions, misinformation, and confusion, and would facilitate

building further efforts such as frameworks, methods, and models for developing and using safe, responsible, and robust AI-based military Cyber Operations.

With a focus on addressing this gap, this research endeavors to achieve a twofold objective. Firstly, to construct a working definition for trustworthy AI within this specific context to serve as a pillar for the second objective, to build a comprehensive framework that will facilitate the development of AI systems within this context, ensuring their adherence to principles of trustworthiness. To attain these objectives, the research employs a transdisciplinary approach by merging insights, techniques, and expertise from multiple domains: AI, military operations, cyber security, and ethics following the Design Science Research methodology and providing exemplifications from military Cyber Operations scenarios.

The outline of this article is structured as follows. Section 2 establishes the background of this research and discusses related studies. Section 3 presents the proposed working definition for trustworthy AI in military Cyber Operations. Section 4 proposes a design framework for building and using trustworthy AI systems in military Cyber Operations with direct exemplifications. Conclusively, Section 5 reflects on the findings of this research and discusses future perspectives.

2. Research Background

The aim of this research is to propose a working definition and framework for building trustworthy AI systems in military Cyber Operations. To reach this goal, a transdisciplinary approach is taken by merging notions and methods from the AI, military operations, cyber security, and ethical domains following the Design Science Research methodology (Peffers et al., 2007; Umbrello & Van de Poel, 2021). Accordingly, the following research questions are defined below:

- How to define the notion of trustworthy AI in the military Cyber Operations context?
- How to design a framework for building and trustworthy AI systems in military Cyber Operations?

Extensive literature review is conducted in the domains above mentioned based on resources found in the ACM Digital Library, IEEE Digital Library, Scopus, and Google Scholar scientific databases using keyword combinations that include trustworthy AI, military operations, and Cyber Operations.

As a starting point, a historical perspective of the trustworthy AI concept is given. Accordingly, Stix (2022) discuss its history and find as main related concepts the following ones: 'ethical AI', 'AI for good', 'beneficial AI', and 'responsible AI'. Moreover, Laux Wachter & Mittelstadt (2023) bring together the notion of TAI (Trustworthy AI) with the one of acceptability arguing that equal attention needs to be provided to both the potential of AI to produce trust to humans and to erode the trustworthiness of humans and public institutions. From a governance perspective, Harrison & Luna-Reyes (2022) discuss the potential of TAI to influence and transform government decision-making processes and threaten democratic values. Along these lines, the EU Commission (2019) propose the following trustworthy AI principles when developing, deploying, and using AI systems: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability. From a legal perspective, Hickman & Petrin (2021) stress that more specificity is necessary in relation to how to harmonize the law rules and governance principles of TAI. Further, the ISO/IEC TR 24028 (2020) standard provides an overview of AI trust issues and analyses the factors that influence the formation of trustworthy AI and its decisions, well-known mechanisms that increase confidence in technical systems, and tackle a series of methods that could contribute to reducing the negative impact on AI credibility.

Focusing on analysing the components and implications of trustworthy AI, Li et al., (2023) introduce a theoretical framework for capturing important aspects of TAI, that include robustness, generalization, explainability, transparency, reproducibility, fairness, privacy preservation, and accountability. The authors consider these aspects in relation to the entire lifecycle of AI systems ranging from data acquisition to model development, deployment, and going to continuous monitoring and governance. Chamola et al., (2023) tackle important aspects necessary for building explainable and trustworthy AI systems and instantiate the framework proposed on autonomous vehicle systems. Moreover, US DHHS (2021) proposed a TAI playbook that captures the building blocks of developing AI systems, considers fair/impartial, transparent/explainable, responsible/accountable, safe/secure, privacy, and robust/reliable as core principles of TAI, and advances a set of considerations and guidelines for implementing these principles at both internal and external levels in each of the phases of AI solutions' development. In relation to the risks and opportunities of AI systems in the military domain, Morgan et al., (2020) define and structure risks in three levels: (i) ethical and legal: LOAC (Law

of Armed Conflict), accountability and moral responsibility, human dignity, human rights, and privacy; (ii) operational: trust and reliability, hacking, poisoning, and adversarial attacks, and accidents and emergent risks; and (iii) strategic: thresholds, escalation management, proliferation, and strategic stability. Yazdanpanah et al., (2021) position responsibility as a fundamental pillar for trustworthy AI for developing autonomous systems.

Flamminiet et al., (2022) define a class diagram for TAI systems that contains as main characteristics robustness (reliability, safety, security, accuracy, usability, interpretability), lawfulness (accountability, privacy, explainability), and ethics (fairness, respect, and transparency). Specifically for Machine Learning (ML) systems, Thuraisingham (2022) considers two important pillars and four layers for building them: security, privacy, integrity, availability as the left pillar, and fairness, unbiased, and anti-discrimination as the right pillar. Li et al., (2022) introduces a framework for trustworthy AI scenarios engineering with the following dimensions: intelligence and index, calibration and certification, and verification and validation. Another perspective is provided by Singh & Singh (2023) which defines AI trustworthiness in relation to the trustee which promotes trustor's goals, provides sound explanations, reasons comprehensively, and models trustor's content (i.e., ability), the trustee promotes the trustor's interests while helping, guiding, and advocating for the trustor (i.e., benevolence), and the trustee promotes the trustor's values, reveals criteria and values honestly, corrects and improves processes, and its help deserves people.

In the cyber security domain, Nassar et al., (2019) propose a framework for building explainable and trustworthy blockchain solutions based on fundamental characteristics: transparency and visibility of transactions, immutability of blocks, traceability and nonrepudiation, and smart contracts. Toussaint & Ding (2020) consider as requirements for building TAI software for IoT systems aspects regarding functionality (e.g., communication, monitorability, performance), business (e.g., policy, quality, regulatory), human (i.e., human factors, usability), trustworthiness (e.g., reliability, resilience, security), timing (e.g., time awareness, time interval and latency), data (e.g., data semantics, data operations, data relations), boundaries (i.e., behavioural, networkability, responsibility), composition (e.g., adaptability, complexity, discoverability), and lifecycle (e.g., deployment, maintainability, productivity). Furthermore, Munir et al., (2023) develop a trustworthy AI mechanism for identifying and explaining in a proactive way cyber risks. In relation to LLMs like ChatGPT, Gupta et al., (2023) consider as a fundamental criterion for assuring the trustworthiness of chatbots and general GenAI systems, data processing correctness. Focusing on the interaction and collaboration between humans and AI systems, Wickramasinghe et al., (2020) propose a set of guidelines for building TAI systems considering the interactions and dynamics between the AI system, system developers, and system users. This perspective is aligned with previous work done for responsible AI-based solutions for military Cyber Operations (Maathuis, 2022). Schlicker & Langer (2021) discuss the actual vs. the perceived trustworthiness: actual trustworthiness is characterized by relevance and availability, while perceived trustworthiness is characterized by detection and utilization.

While this extensive literature review tackles important theoretical and practical notions and methods for building trustworthy AI systems, it reveals the incipient status that these systems have in the ongoing research efforts and discourses in the military and cyber domains. This captures the knowledge gap that the present research aims to tackle.

3. Definition

Trust is a rare concept that "matters for interpersonal relationships, group dynamics, civic engagement, and society at large" expressed in three dimensions: *the how* dimension which captures the psychological foundations of trust; *the whom* which the entity that can be trusted, i.e., a person, group, or beyond humans; and *the what* which captures what is trusted, from simplex meaning a simple aspect or task to multiplex meaning a complex structure of aspects or tasks (Robbins, 2016). Moreover, trust is an underlying function, characteristic, or value that has the power to relegate other values and is based on invisible assumptions (Simpson, 2012). In technology, trust was a central concept regarding technological acceptance and until recent efforts, it was a neglected notion in the AI domain (Renner et al., 2021). Defining and implementing this notion in technological terms is a delicate process. Given recent technological developments in Generative AI, trust in technology became related to trust in AI, i.e., trusting the way AI is built and used, and through this in the underlying AI services and manufactures. This is directly captured in the two main components of trustworthy AI: "(i) its development, deployment, and use should respect fundamental rights and applicable regulation, as well as core principles and values, ensuring an 'ethical purpose', and (ii) it should be technically robust and reliable" (EU Commission, 2019). Stix (2022) identifies five meanings of trustworthy AI: "(i) trust in the proper functioning and safety of the technology, (ii) the technology being worthy of the trust of the

humans making use of it or encountering it otherwise, (iii) humans making use of it or encountering it seeing the technology as trustworthy, (iv) humans making use of it or encountering it experiencing the technology as trustworthy, and (v) the technology that is worthy of trust at all". The human-AI relationship needs to be developed in the paradigm of trust while making sure to (de)construct trustworthiness by embedding corresponding mechanisms that facilitate risks' acceptance that one the one side AI brings (Laux, Wachter & Mittelstadt, 2023) and on the other side humans, the human-AI dynamics considering technical, user, and societal perspectives bring (Laux, Wachter & Mittelstadt, 2023). The essence of the existing trustworthy AI definitions captured based in the extensive literature review conducted in this research is depicted in the table below.

Table 1: Perspectives on Trustworthy AI definition

No	Trustworthy AI Definition	Resource
1	"Trustworthy AI has two components: (1) its development, deployment and use should comply with fundamental rights and applicable regulation as well as respecting core principles and values, ensuring "ethical purpose", and (2) it should be technically robust and reliable."	EU Commission (2019)
2	"Trustworthy AI, in its final form, is defined as being composed of three parts. In order for an AI system to count as "trustworthy," (1) it must be lawful; that is, adhering to all legal obligations which are binding and required at that time, (2) it should be ethical; that is, adhering to and fulfilling all ethical key requirements that have been put forward in the Ethics Guidelines for Trustworthy AI, and (3) it should be robust, both from a technical and a social perspective. The last means that it should be robust in functionality, accurate, reliable, resilient to attack and other cybersecurity and security considerations."	Stix (2022)
3	"Include robustness, security, transparency, fairness, and safety."	Li et al., (2023)
4	"The effort to develop "trustworthy AI" through regulatory laws such as the AI Act acknowledges a need for AI to be trusted if it is to be widely adopted."	Laux, Wachter & Mittelstadt (2023)
5	"All applicable laws and regulations as well as a set of requirements, should be respected by TAI. Specialized evaluation lists attempt to verify the implementation of each of the essential requirements."	Chamola et al., (2023)
6	"Trustworthy AI as programs and systems built to solve problems like a human, which bring benefits and convenience to people with no threat or risk of harm."	Lui et al., (2022)

Conclusively, limited research exists in the military cyber domain, and to the best of our knowledge, this study represents the first initiative in defining trustworthy AI for military Cyber Operations. Hence, given the studies above discussed together with their corresponding definitions, and their transposition in the military cyber domain, and considering the characteristics, limitations, and opportunities of applying AI the military cyber domain, the following definition is proposed:

TAI in MCO = a sub-field of AI that deals with embedding legal, social, and ethical norms, principles, and values in the design, development, deployment, and use of AI systems built and/or used for conducting military Cyber Operations and building and/or using cyber weapons/capabilities.

This definition could be reduced to a series of *functions* that need to be embedded in the *phases* of AI life cycle development when built/used in a specific *context*. These elements are further elaborated:

Functions represented by the legal, social, and ethical norms, principles, and values that need to be implemented in the AI systems used since their design phase given their specific context. Therein, the stakeholders/agents involved deal with a series of challenges at human level (e.g., lack of ethical knowledge and lack of human agency and oversight), technological level (e.g., lack of transparency, clarity, and audit of AI systems, and lack of trust, acceptance, and volatile or buildable risk appetite of decision makers), and context level (e.g., lack of quality data, fairness, and assessment frameworks) (Laux, Wachter & Mittelstadt, 2023).

Phases expressed by the AI life cycle development phases starting from goal definition and design and going to use. This implies a direct commitment to implementing relevant legal, social, and ethical norms like transparency, accountability, and security in every development phase Accordingly, a clear definition is provided in legal, social, and ethical terms and is further translated to corresponding technical terms, methods, and techniques that should be implemented.

Context described by the military Cyber Operations that are independent military operations or part of broader military operations conducted using (intelligent) cyber weapons/capabilities (e.g., different forms of malware such as ransomware and trojans, DDoS attacks) for achieving military goals (Maathuis, Pieters & van

den Berg, 2018a). Examples of such operations are considered to be Stuxnet, the ones conducted in Georgia in 2008, and the ones conducted in Ukraine in 2015, 2016, and since 2022 in the ongoing war. Understanding the context together with its entities, relationships inside as well as outside the context, and the dynamics involved are mandatory for building/using trustworthy AI-based military Cyber Operations.

4. Framework Design

The definition presented blends the functions, phases, and context where the AI systems are built/used when conducting military Cyber Operations, reflecting its socio-technical dimensions and nature. These dimensions (i.e., functions, phases, and context) are further captured in a design framework depicted in Figure 1 and elaborated in Figure 2 in relation to the core trustworthy AI perspectives (i.e., ethical, social, and legal) and the core life cycle phases of AI (i.e., design, development, evaluation, and use). The dimensions together with their corresponding entities and elements are further elaborated and instantiated based on the resources studied in this research.

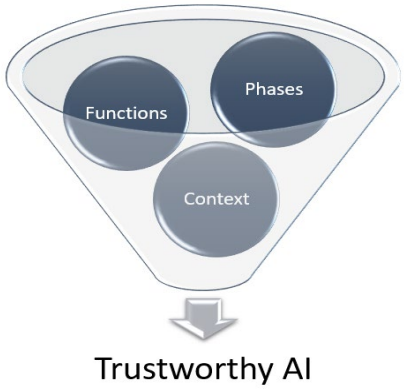

Figure 1: Dimensions of Trustworthy AI in military Cyber Operations

The *Functions* dimension encompasses legal, social, and ethical norms, principles, and values that need to be implemented in the AI systems built/used in military Cyber Operations. These functions are inseparable and are defined as follows:

- *Legal norms* are essential to ensure that conducting a military Cyber Operation is in accordance with established international laws, treaties, and conventions. These norms help regulate the processes and people involved when building and conducting them, assure the minimization of harm to civilians, and uphold fundamental principles of humanity during war. These include IHL (International Humanitarian Law) principles like distinction and proportionality, ROE (Rules of Engagement) that provide specific instructions outlining how to act during an operation, Criminal Law for preventing crimes against humanity or genocide, and Domestic Law of the involved nations.
- *Social principles* contain existing standards and guidelines for building/using AI systems in a trustworthy manner. To recall a few relevant social principles, one could think of the ones proposed by the EU Commission, IEEE, and ISO/IEC standards plus strategies and guidelines proposed by defense organizations.
- *Ethical values* represent fundamental principles that guide the development and use of AI systems in a manner that is morally sound and aligned with societal values. The AI systems should respect human rights, uphold human dignity, and promote fairness. Accordingly, values like transparency, responsibility, security, privacy, robustness, autonomy, fairness, and non-maleficence should be defined and built-in these systems since their design phase.

For exemplification purposes, the focus is on engaging a military target in a joint military Cyber Operation conducted during war by degrading communication lines of the adversary using an intelligent malware (cyber weapon). This directly implies in legal terms to respect the IHL and ROE of the operation, in social terms to be guided by the existing trustworthy AI guidelines and principles, and in ethical terms, to understand and address the core ethical values impact by the development, deployment, and use of this weapon.

The *Phases* dimension captures the life cycle development phases of AI systems developed/used in military Cyber Operations. These phases are:

- In the *Design* phase, the legal norms, social principles, and ethical values of the AI system to be developed need to be in-depth defined and analyzed considering its context, goal, and action. Furthermore, concrete mechanisms need to be defined to address, avoid, and mitigate their expected unintended effects to humans and objects.
- In the *Development* phase, a direct translation is made between the points identified in the previous phase as follows: (i) for transparency, corresponding Explainable AI methods and techniques need to be implemented accounting stakeholders' needs and the explainability-accuracy trade-off; (ii) for responsibility, relevant mechanisms need to be adopted for assuring what is allowed to do and what is right to do; (iii) for security and privacy, considering risk management mechanisms (e.g., vulnerability/impact assessment) and attack prevention (e.g., security and privacy preserving mechanisms) in respect to technical (e.g., data, software) and social (e.g., humans) dimensions; (iv) for robustness, measures that assure the capability and reliability to perform as intended need to be considered focusing on performance, accuracy, and integrity maintenance under various circumstances (e.g., data distribution changes or environmental fluctuations); (v) for autonomy, attention to balancing the decision-making and independence abilities of AI while acknowledging human oversight and control in a sense that the systems should exhibit a proper autonomy level that respects legal, social, and ethical principles; (vi) for fairness, dedicated mechanisms need to be put in place to assure that the AI system is designed, trained, developed, and deployed for avoiding bias, discrimination, and unequal treatment of humans or objects; (vii) for non-maleficence, a clear and responsible approach needs to be adopted for assuring that the AI systems are developed/used according to their goal which should strive to avoid or limit harm, suffering, and damage to humans, objects, and societies at large in an intentional or neglectable way.
- In the *Evaluation* phase, next to assuring that the functionality of the AI systems developed is according to the objectives and requirements defined, a comprehensive approach is crucial for assessing in military, legal, technical, and ethical terms their performance against the pre-defined norms, principles, and values. The evaluation of the systems should be done in real-world mirroring settings and scenarios with well-defined and implemented metrics and mechanisms where the feedback and perspective of directly involved stakeholders should be accounted.
- In the *Use* phase, mechanisms for deployment, monitoring, and use of AI systems should be taken for assuring a continuous and adaptive approach and functionality. Ongoing monitoring is crucial since it implies real-time assessment of the system's performance and compliance with the defined principles. Further, user feedback and experience should be accounted for in relation to the effects produced on direct and collateral entities in the operation.

When building an intelligent cyber weapon, direct measures for embedding all the defined legal, social, and ethical requirements need to be considered in the development phase. For each of these, concrete evaluation methods and metrics need to be included. For instance, including LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive exPlanations) mechanisms for facilitating transparency/explainability of the AI systems developed, and building avoidance or limitation mechanisms for collateral effects produced by the action of these systems on humans and objects for assuring the non-malevolent nature of the AI systems.

The *Context* dimension represents the background where the AI systems are used. Therein, the following entities are found:

- The Operation encompasses the activities and actions taken for achieving military objectives using (intelligent) cyber weapons/capabilities. While these depend on aspects like the mission defined, goals, conflict nature, operational environment, stakeholders, measures for defining, building, and assuring the trustworthiness of the AI systems used need to be applied in all development phases.
- The *Stakeholders* are individuals or groups with distinct roles, interests, and relationships. They either represent the core agents involved when building and conducting military Cyber Operations, or audience that either engages with the core stakeholders or is just impacted by its action as end-users or collateral agents. Through their involvement, the stakeholders are found before-the-loop, in-the-loop, and over-the-loop. The core agents represent the backbone of the development cycle of the AI systems and their duty implies building, deploying, evaluating its performance, continuously upgrading it, and (if necessary) certifying or preparing the systems used for future settings in accordance with legal, military-legal, social, ethical, and military-ethical norms, principles, and values.

- The *Dynamics* refer to relationships between the stakeholders involved/impacted in military Cyber Operations. Stakeholders involved like military Commanders, AI, and cyber engineers interact with the stakeholders impacted, e.g., civilians, neutral and friendly parties, and possible hostile forces either in a direct or indirect way. The dynamics between these stakeholders are key elements that can influence the preparation, execution, action, and impact of military Cyber Operations, and could span dynamics from interdependence, information flow, and stabilization and reconstruction to conflict of interests, and legal and ethical concerns.

For the same scenario, the complexities and uncertainty points need to be considered from the beginning when planning the operation considering the perspectives of involved and impacted stakeholders in an analytical and prescriptive approach. Moreover, the nature and status of the dynamics between the stakeholders involved is of major importance, should be clarified since the beginning of the process, and need to be improved when/where necessary for enhancing trust in humans and the AI systems used.

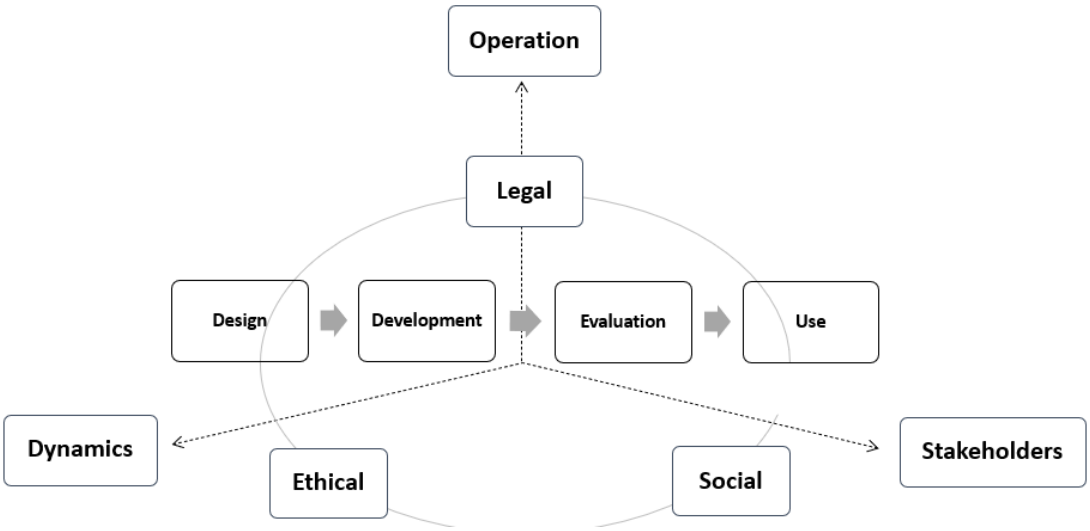

Figure 2: Trustworthy AI Framework in military Cyber Operations

5. Conclusions

In a world increasingly shaped by technology, and in particular AI, a fusion between innovation and trust is required to assure that the intelligent systems built are in accordance with existing legal, social, and ethical norms, principles, and values for becoming safe, responsible, and robust, i.e., trustworthy. While TAI represents the latest AI paradigm, it adopts a socio-technical stance that finds itself in an incipient phase given its complexity and implications now and in the future. In the military cyber domain, dedicated academic and practitioner initiatives and studies for understanding and building trustworthy AI systems are limited, but continue to develop either as a whole or by tackling one or more areas of trustworthy AI, e.g., transparency, robustness, security, and privacy. Nevertheless, a unified effort that defines and bridges the dimensions of trustworthy AI would be beneficial for strengthening awareness, trust, and proper technological adoption, supporting diverse decision-making processes by preventing confusion, dis/misinformation, and strengthening resilience against unexpected events (Fard & Maathuis, 2021). Hence, taking into consideration the identified knowledge gap in this domain, this research aims to formulate a comprehensive definition and build a robust framework for the development and use of trustworthy AI systems in the context of military Cyber Operations. It does that by taking a transdisciplinary approach by merging concepts, methods, and techniques from the AI, military operations, cyber security, and ethics domains following the Design Science Research methodology. For future research perspectives, (i) elaboration of the framework proposed could be considered while focusing on specific phases of building and conducting military Cyber Operations, and (ii) execution of the framework proposed in different types of military Cyber Operations could be regarded. Through these efforts, this research aims to bridge the existing gaps in knowledge and contribute to building the foundation for integrating and using AI in a responsible, safe, effective, and trustable way in military Cyber Operations.

References

Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., & Sikdar, B. (2023). A Review of Trustworthy and Explainable Artificial Intelligence (XAI). *IEEE Access*.

EU Commission (2019). Ethics guidelins for Trustworthy AI.
Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.
Flammini, F., Alcaraz, C., Bellini, E., Marrone, S., Lopez, J., & Bondavalli, A. (2022). Towards trustworthy autonomous systems: Taxonomies and future perspectives. *IEEE Transactions on Emerging Topics in Computing*.
Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access*.
Harrison, T. M., & Luna-Reyes, L. F. (2022). Cultivating trustworthy artificial intelligence in digital government. *Social Science Computer Review*, *40*(2), 494-511.
Hickman, E., & Petrin, M. (2021). Trustworthy AI and corporate governance: the EU's ethics guidelines for trustworthy artificial intelligence from a company law perspective. *European Business Organization Law Review*, *22*, 593-625.
ISO/IEC TR 24028 (2020). Overview of trustworthiness in AI.
Laux, J., Wachter, S., & Mittelstadt, B. (2023). Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*.
Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, *55*(9), 1-46.
Li, X., Ye, P., Li, J., Liu, Z., Cao, L., & Wang, F. Y. (2022). From features engineering to scenarios engineering for trustworthy AI: I&I, C&C, and V&V. *IEEE Intelligent Systems*, *37*(4), 18-26.
Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., ... & Tang, J. (2022). Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, *14*(1), 1-59.
Maathuis, C., Pieters, W., & van den Berg, J. (2018a). Developing a cyber operations computational ontology. *Journal of Information Warfare*, *17*(3), 32-49.
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. (2022). 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. (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).
Morgan et al., (2020). Military applications of Artificial Intelligence. RAND.
Munir, M. S., Shetty, S., & Rawat, D. B. (2023). Trustworthy Artificial Intelligence Framework for Proactive Detection and Risk Explanation of Cyber Attacks in Smart Grid. *arXiv preprint arXiv:2306.07993*.
Nassar, M., Salah, K., ur Rehman, M. H., & Svetinovic, D. (2020). Blockchain for explainable and trustworthy artificial intelligence. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(1), e1340.
Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, *24*(3), 45-77.
Renner, M., Lins, S., Söllner, M., Thiebes, S., & Sunyaev, A. (2021). Achieving Trustworthy Artificial Intelligence: Multi-Source Trust Transfer in Artificial In-Telligence-Capable Technology. In *Forty-Second International Conference on Information Systems, Austin, USA* (pp. 1-17).
Robbins, B. G. (2016). What is trust? A multidisciplinary review, critique, and synthesis. *Sociology compass*, *10*(10), 972-986.
Schlicker, N., & Langer, M. (2021). Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Proceedings of Mensch und Computer 2021* (pp. 325-329).
Simpson, T. W. (2012). What is trust?. *Pacific Philosophical Quarterly*, *93*(4), 550-569.
Singh, A. M., & Singh, M. P. (2023). Wasabi: A conceptual model for trustworthy artificial intelligence. *Computer*, *56*(2), 20-28.
Stix, C. (2022). Artificial intelligence by any other name: a brief history of the conceptualization of "trustworthy artificial intelligence". *Discover Artificial Intelligence*, *2*(1), 26.
Szabadföldi, I. (2021). Artificial intelligence in military application–opportunities and challenges. *Land Forces Academy Review*, *26*(2), 157-165.
Thuraisingham, B. (2022). Trustworthy machine learning. *IEEE Intelligent Systems*, *37*(1), 21-24.
Toussaint, W., & Ding, A. Y. (2020). Machine learning systems in the IoT: Trustworthiness trade-offs for edge intelligence. In *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)* (pp. 177-184). IEEE.
UK DoD (2022). Defence Artificial Intelligence Strategy.
Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283-296.
US DHHS (2020). Trustworthy AI (TAI) Playbook.
Wickramasinghe, C. S., Marino, D. L., Grandio, J., & Manic, M. (2020). Trustworthy AI development guidelines for human system interaction. In *2020 13th International Conference on Human System Interaction (HSI)* (pp. 130-136). IEEE.
Yazdanpanah, V., Gerding, E., Stein, S., Dastani, M., Jonker, C. M., & Norman, T. (2021). Responsibility research for trustworthy autonomous systems.