Proceedings of the
19th International Conference on
Cyber Warfare and Security
University of Johannesburg
South Africa
26-27 March 2024



Dr Jaco du Toit and
Prof Brett van Niekerk

A conference managed by ACI, UK

Proceedings of the

19th International Conference on Cyber Warfare and Security
ICCWS 2024

A Conference
Hosted By

University of Johannesburg
Johannesburg, South Africa

26-27 March 2024

Copyright the authors, 2024. All Rights Reserved.

No reproduction, copy or transmission may be made without written permission from the individual authors.

Review Process
Papers submitted to this conference have been double-blind peer reviewed before final acceptance to the conference. Initially, abstracts were reviewed for relevance and accessibility and successful authors were invited to submit full papers. Many thanks to the reviewers who helped ensure the quality of all the submissions.

Ethics and Publication Malpractice Policy
ACIL adheres to a strict ethics and publication malpractice policy for all publications – details of which can be found here:
http://www.academic-conferences.org/policies/ethics-policy-for-publishing-in-the-conference-proceedings-of-academic-conferences-and-publishing-international-limited/

Self-Archiving and Paper Repositories
We actively encourage authors of papers in ACIL conference proceedings and journals to upload their published papers to university repositories and research bodies such as ResearchGate and Academic.edu. Full reference to the original publication should be provided.

Conference Proceedings
The Conference Proceedings is a book published with an ISBN and ISSN. The proceedings have been submitted to a number of accreditation, citation and indexing bodies including ISI Web of Science and Elsevier Scopus. Author affiliation details in these proceedings have been reproduced as supplied by the authors themselves.

From 2022 these proceedings are open access and freely available for all to read. The Conference Proceedings for previous years can be purchased from **http://academic-bookshop.com**

The Conference Proceedings for this year and previous years can be purchased from
http://academic-bookshop.com

E-Book ISBN: 978-1-914587-97-9
E-Book ISSN: 2048-9889
Print ISBN: 978-1-914587-96-2
Book Version ISSN: 2048-9870

Published by Academic Conferences International Limited

Reading, UK

+44118 324 6938

www.academic-conferences.org
info@academic-conferences.org

Preface **v**

Committee **vi**

Biographies **vii**

Academic Papers

Cybersecurity Implications of Virtual Currency Reward Systems in the Metaverse
Hamza Allimia, Stacey Baror, Hein Venter *1-9*

On The Zero-Trust Intranet Certification Problem
Danielle Botha-Badenhorst, André Martin McDonald, Graham David Barbour, Ethan Buckinjohn, Wian Gertenbach 10-18

Social Media as a Strategic Advantage during Cyberwarfare: A Systematic Literature Review
Errol Baloyi, Oyena Mahlasela, Nokuthaba Siphambili, Mayan Stegmann 19-25

The Role of Industry-Academia Partnerships Can Play in Cybersecurity: Exploring Collaborative Approaches to Address Cybercrime
Benjamin Yankson, Emmanuel Berkoh, Mubarak Hussein, Yvonne Dadson 26-33

Cryptocurrency-crime Investigation: Fraudulent use of Bitcoin in a Divorce Case
Johannes George Botha, Louise Leenen 34-42

Capture the Flag with ChatGPT: Security Testing with AI ChatBots
David Chamberlain, Ellis Casey 43-54

Teenagers: A Social Media Threat Vector
Henry Collier, Charlotte Morton 55-61

Past and Present Russian Information Operations in Ukraine: Competition into Conflict
Stephen Defibaugh 62-65

Ransomware Detection Using Portable Executable Imports
Tanatswa Ruramai Dendere, Avinash Singh 66-74

A Proposed High-Level Methodology on How OSINT is applied in Blockchain Investigations
Wian Gertenbach, Johnny Botha, Louise Leenen 75-83

An Ontology of Cyberspace as a Basis for Decision-making in Cyberoperations
Alexander Grandin 84-94

Learn, Unlearn and Relearn: Adaptive Cybersecurity Culture Model
Tapiwa Gundu 95-102

Anomaly Detection for the MIL-STD-1553B Multiplex Data Bus Using an LSTM Autoencoder
Brian Lachine, Alec Harlow, Vincent Roberge 103-111

Authentication in a Hyperconnected World: Challenges, Opportunities and Approaches
Christoph Lipps, Jan Herbst, Rekha Reddy, Matthias Rüb, Hans D. Schotten 112-120

Human-Centered AI in Military Cyber Operations
Clara Maathuis 121-128

Towards Trustworthy AI-based Military Cyber Operations
Clara Maathuis 129-136

Ukrainian Thoughts and Feelings based on One Year War Content Analysis
Clara Maathuis, Iddo Kerkhof 137-148

Can You Give Me a Lift?
William Mahoney, Adam Spanier 149-155

Cybercrime Classification: A Victimology-based approach
Sayonnha Mandal 156-167

Chain of Custody and Evidence Integrity Verification Using Blockchain Technology
Adir Miller, Avinash Singh 168-176

Cyber Resilience, Dependability and Security
Gareth Davies, Angela Mison, Peter Eden 177-184

Cross-disciplinary AI supply chain risk assessment
Gareth Davies, Angela Mison, Dr. Richard Ward 185-190

Proof of Concept of a Digital Forensic Readiness Cybercrime Language as a Service
Maryam Mohamad Al Mahdi, Stacey Baror 191-199

An Analysis of Cybersecurity Architectures
Noluntu Mpekoa 200-207

Bibliometric Analysis of Cyber Warfare Research in Africa: Landscape and Trends
Jabu Mtsweni, Mphahlela Thaba 208-218

Systematic Review to Propose a Blockchain-based Digital Forensic Ready Internet Voting System
Edmore Muyambo, Stacey Omeleze Baror 19-230

Theory-Guided Feature Selection in Cybercrime Data Science
Shiven Naidoo, Rennie Naidoo 231-239

Infusing Morabaraba game design to develop a cybersecurity awareness game (CyberMoraba)
Mike Wa Nkongolo 240-250

Bahrun Naim's Hacking Manifesto: How A Historical Model of Cyber Threat Mischaracterization Helps Us Diffuse a Dead Cyberterrorist Influencer
Tim Pappa 251-261

Typology of State Actors' Behavior in Cyber Space
Ada Peter, Ujunwa Ohakpougwu 262-267

Friend or Foe – The Impact of ChatGPT on Capture the Flag Competitions
Heloise Pieterse 268-276

Reverse-Engineering of Disinformation Campaigns During the War in Ukraine
Lora Pitman, Ava Baratz, Kelly Morgan, Marcy Alvarado 277-281

Deepfakes: The Legal Implications
Trishana Ramluckan 282-288

Metaverse: Virtual Currencies as a Mechanism for Employee Engagement and Retention
Shelley Robertson, Stacey Baror, Hein Venter 289-298

Cybersafe: Gamifying Cybersecurity Training with a Training App
Carlos Roque, Gareth Moodley, Sayonnha Mandal 299-307

Analysis of Media Influence on Military Decision-Making
Bonnie Rushing 308-316

Building Cybersecurity Capacities in Zambia's Business Sector: Guideline for SMEs
Goni Saar, Rabelani Dagada 317-326

Coming Back Down to Earth: A Grounded Look at Space Cybersecurity in Southeast Asia
Shantanu Sharma 327-337

A Privacy-Compliant Process for Digital Forensics Readiness
Gabriel Shoderu, Stacey Baror, Hein Venter 337-347

Impact of Cyber Security Operations on Hardware Requirements for Stable and Workable Industrial Environments
Jussi Simola, Arttu Takala, Riku Lehkonen, Tapio Frantti, Reijo Savola 348-357

Hardware Sequence Combinators
Stephen Taylor, Gunnar Pope 358-365

A Strategic Path for Digital Transformation in Cyber Warfare for African Militaries
Mphahlela James Thaba, Jabu Mtsweni 366-373

Adversarial Camera Patch: An Effective and Robust Physical-World Attack on Object Detectors
Kalibinuer Tiliwalidi, Bei Hui, Chengyin Hu, Jingjing Ge 374-384

A Case Study on Multi-Countries Money Laundering Scheme and A Proposed Automatic Detection System
Xiao Tan, Tsz-Fung Tony Tse, Siu Ming Yiu, Hiu-Man Human Lam 385-394

Cyber Operations in Peace and War: A Framework for Persistent Engagement
Brett van Niekerk 395-404

Unpacking AI Security Considerations
Namosha Veerasamy, Danielle Badenhorst, Mazwi Ntshangase, Errol Baloyi; Noku Siphambili;
Oyena Mahlasela 405-413

Recognising Cyber Blockades as Crimes Against Humanity: Can International Criminal Law Keep Up?
Dora Vanda Velenczei 414-420

Covert Subversive Agents and Consensus Disruption on Large Projects
Johannes Vorster, Louise Leenen 421-429

Exploring South Africa's Cybersecurity Legal Framework regulating Information Confidentiality, Integrity, and Availability
Murdoch Watney 430-437

PhD Papers

Managing Cyber Security Debt: Strategies for Identification, Prioritisation, and Mitigation
Christo Coetzer, Louise Leenen 439-446

An AI Model for Digital Forensic Readiness in the Cloud using Secure Access Service Edge
taurai Hungwe, Prof Hein Venter 447-454

On the Benefits of Vulnerability Data Consolidation in Application Security
Santanam Kasturi, Xiaolong Li, Peng Li, John Pickard 455-462

Options for Signalling Cyber Deterrence Using Cyber Capabilities
Maria Keinonen, Kimmo Halunen 463-471

A Federated Distributed Digital Forensic Readiness Model for the Cloud
Renico Koen, Hein Venter 472-480

Strengthening Aviation Cybersecurity with Security Operations Centres
Wesley Murisa, Marijke Coetzee 481-489

Quantum-Secure Signalling Model for L1/L2 Next-Gen Interconnect and Roaming Networks Over IPX for NB-IoT Traffic: A Review
Themba Ngobeni, Boniface Kabaso 490-500

Integrating Enterprise Architecture into Cybersecurity Risk Management in Higher Education
Mafika Nkambule, Joey Jansen van Vuuren 501-510

A Study into Privacy and Legal Issues in Cloud Computing: The Mozambican Context
ambrosio Patricio Vumo 511-517

Masters Papers

Using Blockchain to Secure Digital Identity and Privacy Across Digital Sectors
Jaynill Gopal, Stacey Omeleze-Baror 519-526

Defining the "Cyber-Pearl Harbor" - Validation of the DSLP-framework in "Offensive Cyberspace Operations Targeting Ukraine: A Cyber Pearl-Harbor"
Jonathan Lilja, Eleshwa Eishayea, Gazmend Huskaj 527-534

Assessing the Effectiveness of ADS-B Mitigations
Heinke Lubbe, Rudi Serfontein, Marijke Coetzee 535-544

Identifying the Scope of Cybersecurity Research Conducted in the Maritime Industry: 2003 – 2023
Tshepo Mawer, Sune von Solms, Johan Meyer 545-554

National Critical Information Infrastructure Protection through Cybersecurity: A National Government Perspective
Thulisile Dephney Mkhwanazi, Lynn Futcher 555-564

Intelligence Agencies' move to the Cloud: Challenges and Opportunities
Karin Säberg, Gazmend Huskaj 565-573

Enhancing Privacy and Security in Large-Language Models: A Zero-Knowledge Proof Approach
Shridhar Singh 574-582

Securi-Chain: Enhancing Smart Contract Security in Blockchain Systems Through Optimized Access Control
Keanu Swart, Stacey Omeleze Baror, Hein Venter 583-591

Work in Progress Papers

Using Digital Forensics for Android Smartphone Devices to aid Criminal Investigations
Stephanie Agenbag, Andre Henney, Heloise Pieterse 593-596

Deep Learning-based Framework for Detecting Malicious Insider-Inspired Cyberattacks Activities in Organisations
Gibson Chengetanai, Teandai R. Chandigere, Pepukai Chengetanai, Rachna Verma 597-601

Active Gait System for Real-Time Surveillance Against Cyber-Physical Attacks
Glen Moepi, Topside Mathonsi 602-606

Utilization and Sharing of Cyber Threat Intelligence Produced by Open-Source Intelligence
Jyri Rajamäki, Stephen McMenamin 607-611

Implementation of OSINT for Improving an International Finance Sector Organization's Cybersecurity
Jyri Rajamäki, Krista Tiitta 612-616

Art Crime Does not pay: Multiplexed Social Network Analysis in Cultural Heritage Trafficking Forensics
Jarno Salonen, Alessandro Guarino 617-620

ICCWS Preface

These proceedings represent the work of contributors to the 19th International Conference on Cyber Warfare and Security (ICCWS 2024), hosted University of Johannesburg, South Africa on 26-27 March 2024. The Conference Chair is Dr. Jaco du Toit, University of Johannesburg, South Africa, and the Program Chair is Prof Brett van Niekerk, from Durban University of Technology. South Africa.

ICCWS is a well-established event on the academic research calendar and now in its 19th year, the key aim remains the opportunity for participants to share ideas and meet the people who hold them. The scope of papers will ensure an interesting two days. The subjects covered illustrate the wide range of topics that fall into this important and ever-growing area of research.

The opening keynote presentation is given by Prof Joey Jansen van Vuuren, Tshwane University of Technology, City of Johannesburg, Gauteng, South Africa, on the topic of *The threat of ransomware attacks in Higher Education Institutions.* The second day of the conference will open with an address by Dr Manuel Corregedor, Chief Executive Officer at Telspace Africa. His talk is titled *A cyber security tale from the trenches.*

With an initial submission of 163 abstracts, after the double blind, peer review process there are 49 Academic research papers, 8 PhD research papers, 8 Masters Research papers and 6 work-in-progress papers published in these Conference Proceedings. These papers represent research from Australia, Botswana, Canada, China, Finland, France, Germany, Hong Kong, India, Mozambique, Netherlands, New Zealand, Nigeria, Pakistan, Saudi Arabia, Singapore, South Africa, Sweden, Switzerland, UK, and USA.

We hope you enjoy the conference.

Dr. Jaco du Toit
University of Johannesburg
South Africa
March 2024

ICCWS Conference Committee

Dr Bulent Acma, Anadolu University, Eskisehir, Turkey; Dr William ("Joe") Adams, Univ of Michigan/Merit Network, USA; Dr Maamoun Ahmed, Military Technological College, Oman; Prof Hamid Alasadi, Iraq University college, Iraq; Dr Elie Alhajjar, USMA, USA; Prof Todd Andel, University of South Alabama , USA; Prof Antonios Andreatos, Hellenic Air Force Academy, Greece; Leigh Armistead, Oak Ridge National Laboratory, USA; Researcher Jawad Awan, Institute of Information & Communication Technology, Pakistan; Prof Richard Baskerville, Georgia State University, USA; Dr Zakariya Belkhamza, Ahmed Bin Mohammed Military College, Qatar; Prof Vijay Bhuse, Grand Valley State University, USA; Dr Mustafa Canan, Naval Postgraduate School, USA; Dr Kevin Matthe Caramancion, University of Wisconsin, USA; Dr Jim Chen, U.S. National Defense University, USA; Dr Sabarathinam Chockalingam, Institute for Energy Technology, Norway; Dr Nathan Clarke, University of Plymouth, UK; Mr Edwin Covert, Bowhead Speciality Underwriters, USA; Dr Paul Crocker, University of Beira Interior, Portugal; Prof Rabelani Dagada, University of South Africa, South Africa; Evan Dembskey, UNISA, South Africa; Prokopios Drogkaris, University of the Aegean, Greece; Prof Eric Filiol, ENSIBS, Vannes, France & CNAM, Paris, France; Dr Jonathan Fuller, United States Military Academy, USA; Dr Ahmad Ghafarian, University of North Georgia, USA; Dr Scott Graham, Air Force Institute of Technology, USA; Prof Dr Tim Grant, Retired But Active Researcher, The Netherlands; Dr John Gray, Nova Southeastern University, USA; Virginia Greiman, Boston University, USA; Dr Michael Grimaila, Air Force Institute of Technology, USA; Dr George Grispos, University of Nebraska at Omaha, USA; Dr Tapiwa Gundu, Nelson Mandela University, South Africa; Vinay Gupta, Microsoft Corporation, USA; Prof Bill Hutchinson, Edith Cowan University, Australia; Dr Berg Hyacinthe, State University of Haiti, Haiti; Ramkumar Jaganathan, VLB Janakiammal College of Arts and Science (affiliated to Bharathiar University), India; Dr Ezhil Kalaimannan, University of West Florida, USA; Dr Bilge Karabacak, Franklin University, USA; Dr Saltuk Karahan, Old Dominion University, USA; Prof Jesuk Ko, Universidad Mayor de San Andres, Bolivia; Dr Ahmet Koltuksuz, Yasar University, Turkey; Prof Thamizhmaran Krishnamoorthy, Government College of Engineering, Bodinayakkanur, Tamilnadu, India; Dr sanjeev kumar, Career Point University Hamirpur HP India , India; Mr Peter Kunz, Hexagon Geosystems, Germany; Dr Gregory Laidlaw, University of Detroit Mercy, USA; Dr Sylvain (Sly) Leblanc, Royal Military College of Canada, Canada; Louise Leenen, CSIR, Pretoria, South Africa; Prof Martti Lehto, University of Jyväskylä, Finland; Dr Antoine Lemay, CYDEF, Canada; Dr Andrew Liaropoulos, University of Piraeus, Greece; Christoph Lipps, German Research Center for Artificial Intelligence,Germany; Volodymyr Lysenko, University of Washington, USA; Dr Bill Mahoney, University of Nebraska, Omaha, USA; ASSI Haribabu Maruturi, Qiscet, India; Dr Stephanos Mavromoustakos, Indiana Institute of Technology, USA; Dr Paul Maxwell, Army Cyber Institute, USA; Dr Todd McDonald, Air Force Institute of Technology, USA; Dr Terry Merz, Council for Scientific and Industrial Research, USA; Dr Robert Mills, Air Force Institute of Technology, USA; Assc Pardis Moslemzadeh Tehrani, Faculty of Law. University of Malaya, Malaysia; Assc Noluntu Mpekoa, University of Johannesburg, South Africa; Dr Barry Mullins, Air Force Institute of Technology, USA; Prof Antonio Muñoz, University of Málaga, Spain; Dr Funminiyi Olajide, Nottingham Trent University, UK; Prof Abdelnaser Omran, School of Economics, Finance and Banking, Universiti Utara Malaysia, Malaysia; Dr Gilbert Peterson, Air

Force Institute of Technology, Dayton, Ohio, USA; Andy Pettigrew, George Washington University, USA; Dr Jackie Phahlamohlaka, Council for Scientific and Industrial Research, Petoria, South Africa; Ms Heloise Pieterse, CSIR, South Africa; Dr Ajeet Singh Poonia, Bikaner Technical University, India; Prof Carlos Rabadão, Politechnic of Leiria, Portugal; Dr Trishana Ramluckan, University of KwaZulu-Nata, South Africa; Dr Mark Reith, Air Force Institute of Technology, USA; Lt Col Ernest Robinson, U.S. Marine Corps / Air War College, USA; Dr Neil Rowe, US Naval Postgraduate School, Monterey, USA; Prof Lili Saghafi, Canadian International College, Montreal, Canada; Dr Char Sample, US Army Research Laboratory, USA; Prof Leonel Santos, Politechnic of Leiria, Portugal; Ramanamurthy Saripalli, Pragati Engineering College, India; Dr Yilun Shang, Singapore University of Technology and Design, Singapore; Dr Pankaj Kumar Sharma, Government Women Engineering college Ajmer , India; Dr D. Cragin Shelton, Capitol Technology University, USA; Dr Dan Shoemaker, Centre for Assurance Studies, USA; Mr Avinash Singh, University of Pretoria, South Africa; Prof Jill Slay, University of South Australia, Australia; Dr William Spring, University of Hertfordshire, UK; Baror Stacey, University of Pretoria, South Africa; Ms Sneha Sudhakaran, Louisiana State University, USA; Dr Sneha Sudhakaran , Florida Institute of Technology , USA; Dr James Sweeney, Capella University, USA; Mr Unal Tatar, University at Albany - SUNY, USA; Dr Manonmani Thayanithi, Mepco Schlenk Engineering College , India; Dr Le Thi-Thu-Huong, Pusan National University, Korea; Dr Antonio Jorge Tomeu-Hardasmal, University of Cadiz, Spain; Dr Vu Tran, Naval Air Warfare Center - Weapons Division, USA; Dr Eric Trias, Air Force Institute of Technology, USA; Dr Chia-Wen Tsai, Ming Chuan University, Taiwan; Brett van Niekerk, Durban University of Technology, South Africa; Dr Namosha Veerasamy, Council for Scientific and Industrial Research, South Africa; Prof Murdoch Watney, University of Johannesburg, South Africa; Richard Wilson, Towson University, USA; Enes Yurtoglu, Turkish Air War College, Turkey; Dr Zehai Zhou, University of Houston-Downtown, USA; Assc Topside Mathonsi, Tshwane University of Technology, South Africa.

Biographies

Conference and Programme Chairs



Dr Jaco du Toit works as a senior lecturer at the Academy of Computer Science and Software Engineering at the University of Johannesburg. He is the director at the Centre for Cyber Security. Jaco worked for nearly 20 years in the computer industry before becoming a full-time lecturer. His research areas include Cyber Security, focusing on privacy and mobile operating environments. A specific interest to him is research in increasing the protection of private information using decentralised data and access control models.



Prof van Niekerk (PhD, CISM) is an Associate Professor in IT at the Durban University of Technology and a research fellow at the Security Institute for Governance and Leadership in Africa at Stellenbosch University. He is chair of the International Federation of Information Processing (IFIP) Working Group on ICT Uses in Peace and War, and co-Editor-in-Chief of the International Journal of Cyber Warfare and Terrorism. He has participated in working groups for the Paris Call on Trust and Security in Cyberspace, the European Cyber Agora, the international research advisory group for the Global Commission on the Stability of Cyberspace.

Keynote Speakers:



Prof Joey Jansen van Vuuren (PhD) heading the Computer Science Department at Tshwane University of Technology. She was the coordinator of the Cybersecurity Centre of Innovation for the Council for Scientific and Industrial Research (South Africa) with the centre focussing on the promotion of research collaboration, cybersecurity education and the exchange of cyber threats. As the Research Group Leader for Cyber Defence (CSIR) for 9 years, she gave the strategic research direction for the research group that was mainly involved in research for the South African National Defence Force and Government sectors on Cyber Defence. This group has initiated several cybersecurity government initiatives in South Africa. As cyber threats became increasingly important in South Africa with the extensive broadband expansions, she focused her research around cybersecurity and government policy. Her research publications include journal papers, conference papers and book chapters on cybersecurity governance. She has presented on several forums including national and international conferences and has also been invited as keynote speaker at international conferences.



Dr Manuel Corregedor is currently employed as the Chief Executive Officer at Telspace Africa. Manuel has a passion for Information Security and over the years has gained a significant amount of knowledge and experience in the both the technical (operational) and management areas of Information Security. He has also done a significant amount of research in the area of Information Security and holds the following degrees from the University of Johannesburg: a BSc and a BSc Honours in Information Technology, an MSc in Information Technology in the field of Information Security and a PhD Computer Science in the field of Information Security. He also holds a number of Information Security related certifications.

Mini Track Chairs



Dr. Char Sample is a researcher with ICF. Dr. Sample also teaches at Boise State University. She has given guest lecturers at Bournemouth University, Rensselaer Polytechnic University and Royal Holloway University. Dr. Sample has over 20 years' experience in the information security industry. Dr. Sample's research focuses on deception, and the role of cultural values in cybersecurity events. More recently she has begun researching the relationship between human cognition and machines. Presently Dr. Sample is continuing research on modeling cyber behaviors by culture, other areas of research are data resilience, cyber-physical systems and industrial control systems.



Dr. Zakariya Belkhamza is currently a member of faculty at the computer information Systems department, Ahmed Bin Mohammed Military College, Doha, Qatar. His area of expertise is information systems, strategic information systems, system design and architecture. Dr. Zakariya Belkhamza holds a Ph.D. in Information systems. He has published several papers in International journals and has authored and edited several books in the area of information systems.



Dr Stilianos Vidalis, Head of the Networks, Security and Systems subject group at the Department of Computer Science of the University of Hertfordshire. Leading the development of an innovative technology that automates threat assessments. Experienced in all aspects of Information Operations (IO) with excellent understanding of the Information Security (IS) lifecycle. Acted as an instructor for the British Armed Forces Intelligence Personnel on penetration testing and digital forensics for a number of years. Acted as a cyber-expert for the Welsh Government and for the Wales University Officers' Training Corps (WUOTC).



Christoph Lipps, graduated in Electrical and Computer Engineering at the University of Kaiserslautern, where he meanwhile lectures as well. He is a Senior Researcher and Ph.D. candidate at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern, Germany, heading the Cyber Resilience & Security Team of the Intelligent Networks Research Department. He is a member of the committee of the ICCWS and ECCWS, as well as a scientific committee member of various international conferences, and reviewer of several journals. His research focuses on Physical Layer Security (PhySec), Physically Unclonable Functions (PUFs) and Artificial Intelligence (AI).



Dipl.-Phys. **Jan Herbst** graduated with a Physics diploma from the Technical University of Kaiserslautern. He worked in the fields of laser physics, molecular beam epitaxy related to quantum dots, and photoluminescence spectroscopy. As a Researcher and Ph.D. candidate, he is working at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern, Germany. His research includes Biometric Security, entity authentication, (Wireless) Body Area Networks, and Security Applications in Combination with Artificial Intelligence (AI).



Dr Trishana Ramluckan is the Research Manager at Educor Holdings and an Honorary Research Fellow at the University of KwaZulu-Natal's School of Law. Previously she was Postdoctoral Researcher in the School of Law and an Adjunct Lecturer in the Graduate School of Business at the UKZN. She is a member of the IFIP working group on ICT Uses in Peace and War. She holds a Doctor of Administration (IT and Public Governance) and in 2020 she was listed as in the Top 50 Women in Cybersecurity in Africa. Her research areas include Cyber Law and Governance.



Dr. Jason C. Brown is a research scientist and assistant professor at the Army Cyber Institute at West Point. He teaches risk management and organizational security, and as a futurist, he studies emerging threats, technological and social trends, and responses to those threats. A currently serving officer with the U.S. Army, Dr. Brown has worked within the intelligence, information operations, and cyber career fields. He has authored technical reports on the future of extremism, information warfare, cyber enabled financial crimes, and Chinese soft power.



Dr. Jessica Dawson is the Information Warfare Division Chief at the Army Cyber Institute at West Point. She holds a Ph.D. from Duke Sociology and teaches social inequality and social theory. Her research focuses on the digital disruption of social processes. She serves as a special advisor to Army senior leaders on social media threats and information warfare.



Dr. Jessica Dawson is the Information Warfare Division Chief at the Army Cyber Institute at West Point. She holds a Ph.D. from Duke Sociology and teaches social inequality and social theory. Her research focuses on the digital disruption of social processes. She serves as a special advisor to Army senior leaders on social media threats and information warfare.



Christoph Lipps, graduated in Electrical and Computer Engineering at the University of Kaiserslautern, where he meanwhile lectures as well. He is a Senior Researcher and Ph.D. candidate at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern, Germany, heading the Cyber Resilience & Security Team of the Intelligent Networks Research Department. He is not only a member of the committee of the ICCWS and ECCWS, but also scientific committee member of various international conferences, and reviewer of several journals. His research focuses on Physical Layer Security (PhySec), Physically Unclonable Functions (PUFs), Artificial Intelligence (AI), entity authentication, Security in the Sixth Generation (6G) Wireless Systems and all aspects of network and cyber security.



Dr. Char Sample is a researcher with ICF. Dr. Sample also teaches at Boise State University. She has given guest lecturers at Bournemouth University, Rensselaer Polytechnic University and Royal Holloway University. Dr. Sample has over 20 years' experience in the information security industry. Dr. Sample's research focuses on deception, and the role of cultural values in cybersecurity events **Update**. More recently she has begun researching the relationship between human cognition and machines. Presently Dr. Sample is continuing research on modeling cyber behaviors by culture, other areas of research are data resilience, cyber-physical systems and industrial control systems.



Dr. Pardis Moslemzadeh Tehrani, is a Visiting Associate Professor at the Faculty of Law, the University of Malaya. She was a Senior Lecturer at the same university from 2015 to 2022. Her research interests span the area of Information Technology Law, International Humanitarian Law and legal research methodology. Pardis has made scholarly contributions in peer-reviewed journals and presented papers in several national and international conferences and served on many conferences and workshop program committees.



Dr Stilianos Vidalis, Head of the Networks, Security and Systems subject group at the Department of Computer Science of the University of Hertfordshire. Leading the development of an innovative technology that automates threat assessments. Experienced in all aspects of Information Operations (IO) with excellent understanding of the Information Security (IS) lifecycle. Acted as an instructor for the British Armed Forces Intelligence Personnel on penetration testing and digital forensics for a number of years. Acted as a cyber-expert for the Welsh Government and for the Wales

Cybersecurity Implications of Virtual Currency Reward Systems in the Metaverse

Hamza Allimia, Stacey Baror and Hein Venter
University of Pretoria, Hatfield, South Africa

h.allimia@tuks.co.za
stacey.baror@cs.up.ac.za
hventer@cs.up.ac.za

Abstract: In the digital age, the metaverse emerges as a revolutionary platform, intertwining virtual reality, augmented reality, and the internet. Central to its allure is the virtual currency reward system, a dynamic mechanism driving user engagement and economic transactions. However, with innovation comes vulnerability. This paper delves into the pressing question: How do virtual currency reward systems in the metaverse introduce cybersecurity threats, and what measures can safeguard against them? The metaverse's vastness, while offering unparalleled opportunities, is a fertile ground for cyber threats. As users navigate virtual landscapes, engage in transactions, and earn rewards, they become potential targets for cyberattacks. This research, rooted in a comprehensive literature review, identifies the gaps in current cybersecurity measures within the metaverse's virtual currency reward systems. Through a vivid case scenario, we illustrate the real-world ramifications of these vulnerabilities, offering readers a tangible grasp of potential threats. Our methodology, a blend of qualitative analysis and conceptual modelling, dissects the intricate relationship between reward systems and their cybersecurity implications. The findings, derived from rigorous analysis, unveil a set of best practices tailored to combat cybersecurity threats specific to virtual currency reward systems. The distilled insights propose a suite of best practices, encompassing both preventive and reactive strategies tailored for the unique challenges posed by virtual currency systems. This research holds immense value for a diverse audience: metaverse users seeking a secure experience, businesses aiming to establish a foothold in this digital realm, cybersecurity professionals navigating new challenges, and platform developers striving for robustness. In essence, as the metaverse's horizon expands, understanding and fortifying its virtual currency reward systems against cyber threats becomes paramount. This paper offers a roadmap to that secure future, emphasizing the need for vigilance, innovation, and collaboration in the face of evolving cyber challenges.

Keywords: Metaverse, Virtual currency, Reward systems, Cybersecurity, Cyber threats

1. Introduction

In the dynamic digital world, the metaverse, merging virtual reality, augmented reality, and the internet, is redefining human interaction, commerce, and socialization. Central to its evolution is an economic model driven by virtual currency reward systems, as seen in platforms like Decentraland and Sandbox as described by Paulose (2022). These systems, pivotal in shaping the digital economic landscape, are simultaneously introducing new cybersecurity challenges.

The growth of the metaverse brings to light security concerns around virtual currency systems. High-profile incidents such as the Axie Infinity hack reported by Van Boom (2022) and vulnerabilities in Ethereum-based games highlight an array of cyber threats ranging from account intrusions to complex data breaches. This evolving landscape reveals a significant research gap in understanding the specific vulnerabilities of metaverse virtual currency reward systems and their implications for various stakeholders.

Our research addresses this gap through a multi-faceted approach. We blend qualitative analysis with conceptual modelling to dissect the relationship between virtual currency systems and their cybersecurity challenges. Our aim is to identify and analyse inherent vulnerabilities, assess their impacts, and explore potential mitigation strategies. This study is motivated by the crucial role of security in the economic success and user trust within the metaverse. Ensuring the robustness of these virtual currency systems is fundamental to realizing the metaverse's potential as a secure, innovative digital environment.

Following this, Section 2 outlines our methodology, Section 3 reviews existing literature, Section 4 introduces our conceptual model and delves into detailed component analysis. Section 5 presents a practical application, Section 6 provide and in-depth discussion, and Section 7 concludes with key insights and future research directions.

2. Methodology

This section outlines our approach to examining cybersecurity in virtual currency systems within the metaverse. The research problem centres on the emerging cybersecurity vulnerabilities in the metaverse's virtual currency systems, outpacing existing security protocols. Our objective is to understand the interplay between

cybersecurity threats and virtual currency systems, identifying vulnerabilities, assessing impacts, and exploring mitigation strategies.

Our initial step was a targeted literature review on virtual currencies, cybersecurity, and the metaverse's economic aspects, identifying significant knowledge gaps. We developed a conceptual model using a layered approach to map the cybersecurity landscape, encompassing technological, economic, and user dynamics. The proposed solution involves a comprehensive strategy with technological safeguards, regulatory compliance, user awareness, and threat intelligence, aiming for a secure virtual currency environment in the metaverse. Additionally, a hypothetical case scenario provides insights into potential challenges and informs our initial risk mitigation recommendations. This research serves as a foundational exploration for a broad audience, including metaverse users, businesses, cybersecurity experts, and developers, setting the stage for future detailed studies in this evolving field.

3. Overview of Existing Research

The dawn of the Metaverse signifies a new epoch of virtual interactions, with virtual currency reward systems being its linchpin. These systems, while propelling economic activity and user engagement, also become a fertile ground for cybersecurity threats. The literature review traverses through scholarly discourse to aggregate insights on cybersecurity mechanisms protecting these virtual currency reward systems within the Metaverse, setting the stage for the ensuing discussion.

3.1 Blockchain and Security

The discourse on blockchain's capacity to bolster security in the Metaverse is notably highlighted by Ryu et al (2022), proposing a blockchain-based authentication scheme. The exploration of inherent security traits like immutability and transparency by these authors are pivotal for a secure virtual currency reward system environment. Nonetheless, a gap emerges in their discussion regarding blockchain's scalability, especially in high-transaction, densely populated Metaverse realms, necessitating further exploration of scalable blockchain architectures and their Metaverse integration.

3.2 Privacy Preservation

Dastagir et al (2022) propose a smart card-based approach for Non-Fungible Token (NFT) authentication using Non-Interactive Zero Knowledge Proof, marking a stride towards transactional privacy, especially with NFTs. However, the literature reveals a vacuum in discussions around the trade-off between privacy preservation and transaction transparency, calling for a meticulous examination to balance these critical aspects in virtual transactions.

3.3 Economic Dynamics and Security

While Kaur et al (2023) delve into the economic dynamics of virtual currencies in the Metaverse, a lacuna emerges regarding the cybersecurity implications, hence, linking economic theories with cybersecurity frameworks remains a pressing need. Similarly, the discussions by Zimmermann (2019) and Mikołajewicz-Woźniak et al (2015) on monetary policies, though hinting at security implications, lack a direct exploration of harmonizing these policies with cybersecurity frameworks to safeguard virtual assets.

3.4 Reward Systems and Security

The works of Pascheka and Düring (2015) and Zhang et al (2022) offer insights into cooperative decentralized decision-making and social patterns within reward mechanisms, yet a deeper analysis on integrating security protocols within reward systems and their real-world resilience in the Metaverse is warranted.

3.5 Threat Landscape and Countermeasures

Vadlamudi (2022) systematizes Metaverse security challenges and countermeasures, providing a structured threat landscape understanding, yet an extension analysing the efficacy and scalability of these countermeasures in virtual currency reward systems is desired. Chen et al (2022) and Huang et al (2023) survey Metaverse security and privacy, establishing a foundational understanding, yet could delve further into specifics regarding virtual currency reward systems to provide a more focused insight into apt security frameworks for these systems.

3.6 Social Engineering and User Behaviour

Deng et al (2023) explore social engineering threats, emphasizing the importance of understanding human-centric vectors of cybersecurity threats. Yet, extending the scope to elucidate the impact on virtual currency reward systems and proposing user-centric security measures is essential.

3.7 Future Avenues and Emerging Threats

Di Pietro and Cresci (2021) alongside Conti et al (2018) broach emerging security and privacy issues, opening avenues for novel threat exploration. However, a more tailored understanding of how these threats could jeopardize virtual currency reward systems is needed.

The review highlights blockchain technology's potential and challenges in Metaverse security, particularly regarding scalability and balancing privacy with transparency. These insights direct our next step: In Section 4, we introduce a conceptual model specifically crafted to address these gaps. This model, stemming from our comprehensive analysis, aims to enhance the security of virtual currency reward systems in the Metaverse, offering a strategic approach to navigate these complexities.

4. Virtual Currency Rewards Model

This section presents a conceptual model designed to explore the intricate facets of virtual currency reward systems within the metaverse, considering the rapid evolution of technology and increasing sophistication of cyber threats. We explore each layer of our conceptual model, focusing on their individual and collective impact on the cybersecurity of virtual currency systems in the metaverse.

The model adopts a layered approach that integrates technological architecture, cybersecurity threats, economic implications, and user dynamics, aiming to serve as a foundation for both academic research and practical applications in the evolving metaverse field. It comprises eight interrelated layers, depicted in Figure 1. The model starts with the 'Rewards System' layer, which drives user engagement, supported by the 'Technological Architecture' layer that ensures system functionality and scalability. Critical 'Cybersecurity Threats' and 'Data Flow and Protection' layers are then examined for their role in maintaining user trust and system integrity. Influencing these is the 'User Behaviour, Trust, and Usability' layer, shaped within the 'Regulatory and Policy Landscape'. The model also considers the 'Economic Implications' of the metaverse's economy, grounded in 'Ethical Considerations' to uphold moral standards across functionalities.

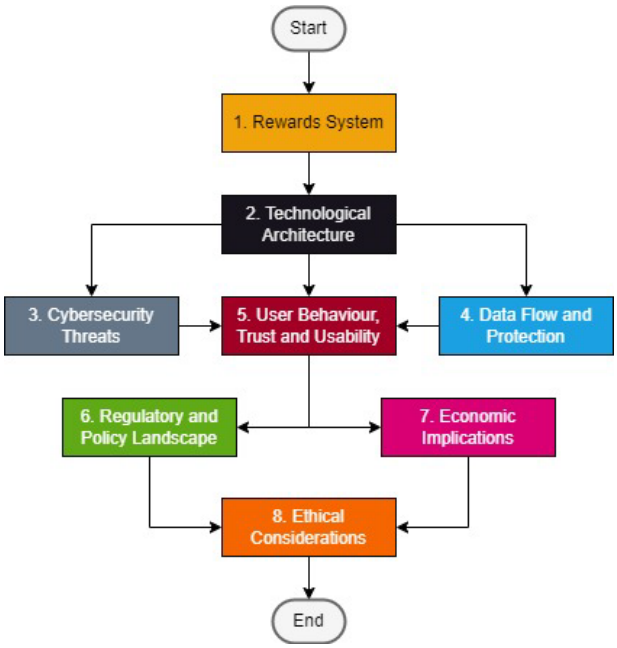


Figure 1: A visual representation of the conceptual model, clearly depicting the interconnectedness of the different layers

4.1 Rewards System

This is the economic engine of the metaverse, akin to systems found in platforms like Decentraland and The Sandbox. It facilitates dynamic interactions through earning and spending mechanisms, shaping user behaviour, and reflecting the economic vitality of the metaverse. The system gamifies user engagement through tasks and achievements, like those seen in real-world behavioural studies on gamification and user engagement. Management tools like advanced virtual wallets are crucial for secure and efficient handling of virtual currency, ensuring a robust economic ecosystem.

In the metaverse, user engagement is incentivized through gamified experiences, paralleling real-world loyalty programs. Users earn virtual currency for activities ranging from daily logins to complex challenges. For example, platforms like Roblox use these mechanisms to enhance engagement, much like frequent flyer programs in airlines. Future updates may introduce AI for personalized experiences.

Spending mechanisms allow users to utilize their earned virtual currency, influencing its perceived value and utility. This mirrors consumer spending behaviour in real economies, where choices impact market dynamics. In Second Life, for instance, users spend on virtual goods and services, enriching their experience and driving platform loyalty. Evolutions in virtual marketplaces are expected.

Effective management tools are exemplified by advanced virtual wallets, which include features like multi-factor authentication and blockchain integration, much like Ethereum's wallet system. These tools provide security through encryption and data loss prevention, enhancing user trust and transaction transparency.

4.2 Technological Architecture

This layer forms the structural backbone of the Metaverse, integrating AI and machine learning to bolster transactional security and enhance user experience. Blockchain technology is central to transactional security, while cloud storage and distributed computing platforms provide essential data management and computational efficiency, with a focus on robust encryption and access control.

Blockchain technologies in virtual currency systems, such as those used in Ethereum, enhance transaction verification security and transparency. This technology operates on a network of nodes, reducing risks of system failure and fraudulent activities and ensuring transparent and immutable transaction records. With future strategies including quantum-resistant encryption against quantum computing threats.

Cloud storage, exemplified by services like Amazon Web Services, plays a critical role in system scalability and accessibility. It offers scalable data storage with strong encryption algorithms for user data confidentiality and robust access control mechanisms to prevent unauthorized access.

Distributed computing platforms, like those provided by Google Cloud, maintain system performance by distributing computational tasks across multiple machines. They exhibit resilience to failures, ensuring uninterrupted service, and maintain data consistency across nodes, crucial for the integrity of virtual currency systems.

4.3 Cybersecurity Threats

This layer acts as a defence against evolving cybersecurity threats, upholding principles like integrity, availability, and confidentiality. It covers recent examples of cybersecurity breaches and the necessary proactive measures. Emphasis is placed on the continuous monitoring and updating of security protocols to protect the virtual currency reward system.

Blockchain, while robust, has vulnerabilities such as the 51% Attack, which allows manipulation of transaction history. This vulnerability, as seen in some cryptocurrencies like Bitcoin, is mitigated by resistant consensus algorithms and regular audits of smart contracts.

The user interface, as the initial interaction point, must be secured against phishing and clickjacking attacks. For example, platforms like Facebook have implemented two-factor authentication and user education strategies to mitigate these risks.

Databases, storing sensitive data, must be protected against SQL injection, data breaches, and DoS attacks. Large databases like Oracle employ encryption, regular audits, and strong access control policies to safeguard against these threats.

4.4 Data Flow and Protection

Focusing on the data lifecycle, this layer addresses challenges and opportunities in data management. It includes case studies showcasing best practices and discusses the influence of emerging data protection technologies and regulations on data management strategies.

Data generation in the metaverse, like practices in financial institutions, involves secure data capture at the user interface. This includes ensuring input validation and initial encryption to secure user-provided data like transaction details.

Post-generation, data is stored in databases or cloud storage, each with unique security considerations. Relational databases, used for structured data, need protection against threats like SQL injections, while cloud storage offers scalability but with risks regarding data sovereignty and unauthorized access. Encrypting data at rest, regular backups, and strong access control policies are pivotal for securing stored data.

Data transmission in virtual currency systems requires secure protocols like SSL/TLS, like those used in online banking systems, to protect data during transit. Additional measures like data masking and intrusion detection systems are crucial to maintain data confidentiality.

Data processing involves secure handling of transaction data, akin to techniques used in high-frequency trading platforms. This includes batch and real-time processing with robust security measures to prevent fraud and ensure efficient data management.

4.5 User Behaviour, Trust, and Usability

Exploring psychological and behavioural aspects, this layer examines how design elements affect user perception, trust, and security awareness. It draws on user experience research specific to virtual environments, highlighting the significance of usability in influencing security behaviour.

User experience is vital for user engagement and the economic viability of virtual currency systems. Key factors include ease of use, which encourages user activity, though overly simplified interfaces may compromise security. Feedback mechanisms like notifications or alerts improve user experience and security. Transaction processing speed affects user satisfaction, with slow systems potentially driving users towards less secure alternatives.

Interface design influences user behaviour and security choices. For example, Apple’s intuitive design principles guide users towards secure actions, balancing usability with robust security features.

User trust is influenced by transparency and past security incidents. Twitter's approach to user data transparency serves as an example of how openness about security measures can enhance user trust.

4.6 Regulatory and Policy Landscape

‘Regulatory and Policy Landscape’ explores international regulations, platform-specific policies, and challenges posed by the metaverse's borderless nature. It includes a detailed analysis of international regulations like GDPR and examines the challenges in forming a unified regulatory framework for the metaverse.

International regulations like Anti-Money Laundering (AML), Know Your Customer (KYC) and GDPR (General Data Protection Regulation) significantly impact virtual currency systems. The implementation of GDPR in EU-based platforms illustrates the complexities of complying with data protection laws in a global context.

Platforms set internal rules to govern interactions and security. LinkedIn's user agreements and community guidelines exemplify how platforms can design policies to maintain security and govern user behaviour.

The borderless nature of the metaverse poses challenges in legal jurisdiction, as seen in global e-commerce platforms like eBay. This complexity underscores the importance of understanding legal frameworks for compliance and user trust.

4.7 Economic Implications

This layer examines the economic impact of virtual currency reward systems and their security measures on the metaverse and its stakeholders. It considers aspects such as monetary policy, inflation, taxation, market efficiency, and user behaviour. It also explores the trade-offs and challenges involved in designing and implementing effective economic models for virtual currency reward systems.

Market dynamics explores how user base growth and scarcity mechanisms, illustrated by Bitcoin, affect virtual currency value, considering integration with traditional markets and the rise of decentralized finance (DeFi). The integration of virtual currencies into traditional financial markets, as seen with Bitcoin, shows potential impacts on global financial systems, influencing investment behaviours and monetary policies.

Economic incentives in platforms like Uber, through referral programs, illustrate how user engagement and virtual currency demand can be boosted, enhancing platform revenue and economic viability.

4.8 Ethical Considerations

This layer delves into the ethical implications of virtual currency reward systems and their security measures on the metaverse and its stakeholders. It explores aspects such as protecting user privacy, ensuring data protection, obtaining user consent, ensuring transparency and accountability, and promoting fairness. It also analyses the trade-offs and challenges involved in designing and implementing ethical frameworks for virtual currency reward systems.

Ethical handling of user data in virtual currency systems, exemplified by Apple's data protection policies, is crucial for maintaining user trust and system integrity. Ethical data handling and privacy are prioritized in accordance with industry best practices, such as those followed by Apple. Additionally, the model is designed to anticipate and address ethical challenges that may arise from advancements in AI technology. The model also advocates for responsible vulnerability disclosure practices, as seen in Microsoft's approach with Coordinated Vulnerability Disclosure (CVD), to maintain system security and user trust.

This model, with its comprehensive approach to virtual currency systems in the metaverse, sets the foundation for a practical exploration in the next section. While it adeptly combines technical and user-focused elements, its current scope suggests room for refinement, especially in its application across various metaverse platforms. Acknowledging this, we next present a case scenario in Section 5 that applies our model in a specific metaverse environment.

5. Case Scenario

In the digital realm of MetaSphere, the virtual economy thrives on the MetaReward system, a sophisticated virtual currency rewards platform. As MetaSphere's user base burgeons, the intricacies of MetaReward's cybersecurity framework come under scrutiny.

5.1 Blockchain Vulnerabilities

MetaReward's reliance on blockchain technology for transaction verification and smart contract executions faces a significant threat when a '51% Attack' is orchestrated by a rogue entity, ShadowMiners. This attack, exploiting a weakness in the network's consensus algorithm, leads to double spending of MetaCoins, devaluing the currency, and undermining trust in the blockchain infrastructure. A detailed analysis reveals that ShadowMiners amassed computational power by infiltrating multiple nodes. The response involves an emergency deployment of a more resilient consensus algorithm and a temporary suspension of transactions to mitigate the attack's impact.

5.2 User Interface, Trust, and Regulatory Adherence

As MetaSphere's user base grows, its user interface reveals vulnerabilities. A sophisticated phishing scheme, engineered to mimic the MetaReward interface, deceives users into divulging their credentials. This incident highlights the need for robust multi-factor authentication and ongoing user education programs about cybersecurity risks. Additionally, a data breach, resulting from an overlooked vulnerability in user data storage, further diminishes trust. This breach prompts a thorough review of data security protocols and a public transparency campaign to restore user confidence. Regulatory scrutiny following these incidents leads to the adoption of stricter compliance measures, aligning MetaSphere's operations with international cybersecurity standards.

5.3 Database Threats

An SQL injection attack targets MetaSphere's database, leaking sensitive transaction histories and user data. This breach, combined with vulnerabilities in data transmission protocols exploited during a high-volume trading event, causes financial discrepancies in MetaCoin balances. The response includes the implementation of robust data encryption, stringent access control mechanisms, and real-time monitoring to secure user data.

Additionally, the introduction of secure data transmission channels like SSL/TLS protocols and enhanced monitoring systems helps prevent similar incidents in the future.

5.4 Economic Implications and Ethical Considerations

The orchestrated attacks disrupt MetaReward's market dynamics, affecting MetaCoins' value. This situation leads to a broader discussion on the economic effects of cybersecurity in virtual currency systems. Ethical dilemmas arise regarding user data protection and privacy, underscoring the importance of maintaining ethical practices to secure the digital economy.

5.5 Future Proofing

These events underscore the need to future proof MetaReward against advancing cybersecurity threats. Strategies include adopting adaptive cybersecurity frameworks, continuous monitoring, and incorporating emerging technologies like AI for predictive threat analysis and quantum cryptography for enhanced data security. This approach aims to ensure the long-term security and viability of virtual currency systems in the metaverse, fostering a dialogue for robust cybersecurity measures and a secure, economically sustainable metaverse ecosystem.

The MetaSphere case scenario underscores the cybersecurity challenges in virtual environments. Moving into Section 6, we expand on these insights, discussing broader implications and mitigation strategies for secure virtual currency systems in the Metaverse.

6. Discussion

This section synthesizes our comprehensive exploration of cybersecurity in virtual currency reward systems within the Metaverse, providing insights and bridging identified gaps. We aim to contextualise our findings within the broader cybersecurity and economic discourse, considering the dynamic nature of the Metaverse.

6.1 Taxonomy of Threats

Our inquiry has delineated a taxonomy of threats and vulnerabilities pertinent to virtual currency reward systems in the Metaverse:

- Regulatory Oversight and Operational Risks: The lack of centralized authority in the Metaverse poses risks for fraud, manipulation, and operational mishaps, undermining economic stability and user trust.
- Smart Contract Vulnerabilities and Integration Risks: Flaws in smart contracts and vulnerabilities in third-party integrations can lead to unauthorized access, system compromises, and potential manipulation of contract outcomes.
- User Vulnerability to Phishing and Social Engineering: The novelty of the Metaverse and lack of user awareness can lead to susceptibility to phishing and social engineering attacks, risking sensitive information and asset security.
- Data Privacy and Asset Security: The extensive collection of user data and in-game assets in the Metaverse, if not secured, can be exploited, leading to privacy breaches and asset theft or manipulation.
- Money Laundering and Market Manipulation: The decentralized nature of virtual currencies can facilitate money laundering and market manipulation, impacting the transparency and integrity of financial transactions.
- Exchange Rate Volatility and Economic Instability: High volatility in virtual currency values can affect transaction utility and stability, leading to economic uncertainty within the Metaverse.
- Exploitation of Reward Mechanisms and Lack of Transparency: Malicious exploitation of reward mechanisms and lack of transparency in reward algorithms can distort economic balance, erode trust, and deter user engagement.
- Account Security Threats: The accumulation of rewards and virtual currency makes user accounts targets for cybercriminals, leading to potential account takeovers and asset loss.
- Asset Vulnerabilities and Rewards Theft: Vulnerabilities in in-game assets and rewards, including theft, duplication, or unauthorized modification, can undermine the Metaverse's economic stability and user trust.
- Comprehensive Security Challenges: Addressing these threats requires a holistic approach, encompassing enhanced security measures, user education, and robust regulatory frameworks to ensure a secure and stable Metaverse environment.

6.2 Proposed Countermeasures

Considering the identified threats, a collection of countermeasures has been postulated with the aim of mitigating the inherent risks associated with virtual currency reward systems:

- Enhanced Security Measures: Alongside technical measures like Multi-Signature Wallets and Multi-factor Authentication, we propose the integration of organizational security cultures and regular cybersecurity training for staff involved in the maintenance of these systems.
- Routine Security Auditing and Real-time Monitoring: We suggest regular security audits, inspired by the practices in financial institutions, and continuous monitoring, leveraging AI and machine learning for detecting suspicious patterns and anomalies.
- Educational Initiatives: Implementing comprehensive User Education Programs, like those used in online banking, can enhance user ability to recognize and avert potential threats like phishing.
- Regulatory and Policy Implementation: Establishing regulatory frameworks should involve collaboration with existing financial and internet governance bodies to create adaptable and enforceable regulations. Transparent reward algorithms, modelled after successful online loyalty programs, can foster user trust while deterring malicious activities.
- Blockchain Utilization: We recommend the use of blockchain not just for identity verification but also for transparent record-keeping and audit trails, enhancing trust and accountability.

6.3 Addressing Complex Challenges

- Operational Risks and Market Manipulation: We delve deeper into managing operational risks by looking at risk management strategies employed in online trading platforms. To combat market manipulation, we propose monitoring mechanisms like those used by stock exchanges.
- Scalability and Evolution of Threats: As the Metaverse expands, solutions must be scalable. We discuss how emerging technologies like cloud computing and edge computing can be leveraged to scale security measures. Additionally, we emphasize the need for adaptive security strategies that evolve with emerging threats, drawing parallels to how cybersecurity measures have evolved in the digital banking sector.
- Implementation Challenges: We acknowledge the complexity of implementing these solutions in a decentralized environment. We propose a phased implementation strategy and highlight the need for collaboration between different stakeholders in the Metaverse.

Our approach combines technical solutions with organizational and regulatory strategies to adapt to the Metaverse's evolving landscape, offering a practical roadmap for securing virtual currency systems. In the next section, we provide a summary of our key findings and propose directions for future research in this dynamic field.

7. Conclusions

Our study on virtual currency reward systems in the evolving Metaverse has revealed complex challenges in cybersecurity, technology, economics, and ethics. We have dissected the Metaverse's economic framework, highlighting the role of virtual currencies and examining the technological infrastructure, including blockchain, cloud storage, and distributed computing, to protect these systems. Our analysis of a wide range of cybersecurity threats underscores the need for robust protective measures and the importance of user experience and trust in the security and economic health of these systems.

We emphasize the necessity for harmonized regulatory approaches in the borderless Metaverse and offer insights into market dynamics and financial integrity. To connect theory with practice, we recommend using practical examples from existing digital economies and focusing future research on emerging technologies, evolving threats, and regulatory implementation in the Metaverse.

As the Metaverse grows, we highlight the need for scalable and adaptable security solutions and interdisciplinary collaboration across various fields to address these multifaceted challenges. Recognizing the nascent state of the Metaverse, we call for continuous research to refine and evolve our solutions in line with the Metaverse's development, aiming for a secure, transparent, and economically robust virtual world.

References

Aks, S. M. Y., Karmila, M., Givan, B., Hendratna, G., Setiawan, H. S., Putra, A. S., Winarno, S. H., Kurniawan, T. A., Simorangkir, Y. N., Taufiq, R., Herawaty, M. T., & Asep (2022). "A Review of Blockchain for Security Data Privacy with

Metaverse", in *2022 International Conference on ICT for Smart Society (ICISS)*, pp. 1-5. https://doi.org/10.1109/ICISS55894.2022.9915055

Chen, Z., Wu, J., Gan, W., & Qi, Z. (2022). "Metaverse Security and Privacy: An Overview", in *2022 IEEE International Conference on Big Data (Big Data)*, pp. 2950-2959. https://doi.org/10.1109/BigData55660.2022.10021112

Conti, M., Kumar, E. S., Lal, C., & Ruj, S. (2018). "A Survey on Security and Privacy Issues of Bitcoin", *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 3416-3452. https://doi.org/10.1109/COMST.2018.2842460

Dastagir, M. B. A., Tariq, O., & Han, D. (2022). "A Smart Card based Approach for Privacy Preservation Authentication of Non-Fungible Token using Non-Interactive Zero Knowledge Proof", in *2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*, pp. 2428-2435. https://doi.org/10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00339

Deng, M., Zhai, H., & Yang, K. (2023). "Social engineering in metaverse environment", in *2023 IEEE 10th International Conference on Cyber Security and Cloud Computing (CSCloud)/2023 IEEE 9th International Conference on Edge Computing and Scalable Cloud (EdgeCom)*, Jul. 2023, pp. 150-154. https://doi.org/10.1109/CSCloud-EdgeCom58631.2023.00034

Di Pietro, R. & Cresci, S. (2021). "Metaverse: Security and Privacy Issues", in *2021 Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pp. 281-288. https://doi.org/10.1109/TPSISA52974.2021.00032

Gadekallu, T. R., Huynh-The, T., Wang, W., Yenduri, G., Ranaweera, P., Pham, Q.-V., da Costa, D. B., & Liyanage, M. (2022). "Blockchain for the Metaverse: A Review", *arXiv:2203.09738 [cs.SI]*. https://doi.org/10.48550/arXiv.2203.09738

Huang, Y., Li, Y. J., & Cai, Z. (2023). "Security and Privacy in Metaverse: A Comprehensive Survey", *Big Data Mining and Analytics*, vol. 6, no. 2, pp. 234-247. https://doi.org/10.26599/BDMA.2022.9020047

Kaur, N., Saha, S., Agarwal, V., & Gulati, S. (2023). "Metaverse and Fintech: Pathway for Innovation and Development", in *2023 3rd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, pp. 1-6. https://doi.org/10.1109/ICIPTM57143.2023.10117956

Mikołajewicz-Woźniak, A. & Scheibe, A. (2015). "Virtual currency schemes – the future of financial services", *Foresight*, vol. 17, no. 4, pp. 365-377. https://doi.org/10.1108/FS-04-2014-0021

Pascheka, P. & Düring, M. (2015). "Advanced cooperative decentralized decision making using a cooperative reward system", in *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, pp. 1-7. https://doi.org/10.1109/INISTA.2015.7276779

Paulose, S. (2022). "Top 5 Metaverse Games to Earn Crypto Rewards in 2022", [online] LinkedIn, www.linkedin.com/pulse/top-5-metaverse-games-earn-crypto-rewards-2022-seby-paulose.

Ryu, J., Son, S., Lee, J., Park, Y., & Park, Y. (2022). "Design of Secure Mutual Authentication Scheme for Metaverse Environments Using Blockchain", *IEEE Access*, vol. 10, pp. 98944-98958. https://doi.org/10.1109/ACCESS.2022.3206457

Vadlamudi, S. (2022). "The Taxonomy of Security issues and Countermeasures in the Metaverse World", in *2022 International Conference on Recent Trends in Microelectronics, Automation, Computing and Communications Systems (ICMACC)*, pp. 553-558. https://doi.org/10.1109/ICMACC54824.2022.10093534

Van Boom, D. (2022). "A Fake Job Offer Reportedly Led to Axie Infinity's $600M Hack", [online], CNET, www.cnet.com/personal-finance/crypto/a-fake-job-offer-reportedly-led-to-axie-infinitys-600m-hack.

Wang, Y., Su, Z., Zhang, N., Xing, R., Liu, D., Luan, T. H., & Shen, X. (2023). "A Survey on Metaverse: Fundamentals, Security, and Privacy", *IEEE Communications Surveys & Tutorials*, vol. 25, no. 1, pp. 319-352. https://doi.org/10.1109/comst.2022.3202047

Zhang, G., Wu, J., Jeon, G., Chen, Y., Wang, Y., & Tan, M. (2022). "Towards understanding metaverse engagement via social patterns and reward mechanism: A case study of nova empire", *IEEE Transactions on Computational Social Systems*, pp. 1-12. https://doi.org/10.1109/TCSS.2022.3211679

Zimmermann, C. D. (2019). "Monetary policy in the digital age", in *Oxford University Press eBooks*, pp. 99-111.

On the Zero-Trust Intranet Certification Problem

Danielle Badenhorst, Graham Barbour, André McDonald, Wian Gertenbach and Ethan Buckinjohn
Council for Scientific and Industrial Research, Pretoria, South Africa

DBadenhorst@csir.co.za
GBarbour@csir.co.za
AMcdonald@csir.co.za

Abstract: Securing corporate networks and ensuring the trustworthiness of network resources are critical security concerns for organisations in today's interconnected digital landscape. The *zero-trust security model* is an approach to designing and implementing ICT systems which prescribes that clients and servers cannot be trusted automatically, even when connected to networks traditionally considered trusted. The implementation of the zero-trust model within the corporate *intranet* requires a secure method to verify the identity of local servers. On the *Internet*, trust in the identity of public servers is established by well-known public Certificate Authorities (CAs), which issue digital certificates to securely identify servers. However, local *intranet* servers exist within the internal address space of the network. Consequently, it is impossible to naturally obtain digital certificates for these servers, validly signed by a public CA, without publicly disclosing sensitive information such as intranet server Domain Name System (DNS) records. This leaves organisations with the option of relying on endpoint management systems to install custom CA root certificates on all corporate browsers or, in some cases, ignoring the problem altogether. In this paper, we draw on practical experience in the deployment of cybersecurity devices in corporate intranets to formally define the *intranet certification problem*. We specify five requirements that a solution to this problem must satisfy. We then conduct a comprehensive review of existing candidate solutions and academic research relevant to the intranet certification problem. Specifically, existing ICT systems for public key infrastructure and endpoint management are identified and evaluated with respect to their ability to meet the stated requirements for solving the intranet certification problem, as well as their cost. Our study reveals that solutions that meet the technical and security requirements of the intranet certification problem are beyond the reach of smaller private sector companies and public sector organisations in underdeveloped and emerging economies. The high cost and technical expertise required for their implementation and management render these solutions impractical. Consequently, by relying on servers with self-signed certificates, these entities inadvertently leave their servers susceptible to impersonation, information theft, and unauthorised resource access, thus violating the fundamental principles of the zero-trust model. We conclude that a gap exists for a simple, cost-effective, and easily managed solution to the intranet certification problem.

Keywords: Zero-trust, Public-key infrastructure, Endpoint management, Man-in-the-middle, Self-signed certificates, Intranet

1. Introduction

Network perimeters have become increasingly blurred with the advent of cloud computing and the widespread use of virtual private networking. Securing corporate networks and ensuring the trustworthiness of network resources in this interconnected landscape are critical security concerns (He et al., 2022). The zero-trust security model offers a promising approach to address these challenges, advocating for continuous entity verification and strict access control (Bobbert and Scheerder, 2022).

The *zero-trust security model*, also referred to as the *zero-trust architecture*, describes an approach to the design and implementation of Information and Communication Technology (ICT) systems (Sarkar et al., 2022; Stankard, 2021). The zero-trust model assumes that there is no traditional network perimeter; networks can be located in the cloud, locally, or as a hybrid of both. Hence, the model prescribes that clients and servers cannot be trusted automatically, even if they are connected to permissioned networks that have traditionally been considered trusted environments. The zero-trust model is implemented by establishing strong identity verification, validating device compliance before granting access, and ensuring least privilege access to only explicitly authorised resources. Zero-trust prescribes a "first verify, *and then* trust" approach (Bobbert and Scheerder, 2022; He et al., 2022). This differs from perimeter-based security, which follows a "*trust and verify*" approach.

The need to implement the zero-trust security model in the corporate intranet is attributed to two factors that erode trust within these networks (Stankard, 2021):

- Modern corporate networks consist of many interconnected zones, cloud services and infrastructure, connections to remote and mobile environments, and connections to nonconventional IT, such as Internet of Things (IoT) devices. This creates greater opportunity for network access by unauthorised entities.

- Modern cyberattacks readily bypass the network boundary and technologies such as firewalls. Examples of such attacks include phishing, social engineering, and a disgruntled employee with access to the corporate intranet.

The zero-trust model is not widely implemented in practice at present. A 2023 survey by Fortinet indicated that only 23% of respondents believed that their zero-trust strategy was fully implemented, citing a lack of information on how to best select zero-trust solutions as the primary challenge (Fortinet, 2023). However, the zero-trust model is gaining recognition as an important paradigm shift in the ICT community, and its acceptance and adoption is gaining momentum in a post-pandemic world. This is due, in part, to the issuing of a presidential executive order in the United States in May 2021 (Young, 2022). The order mandates all federal agencies to adopt zero-trust, providing a timeline until autumn 2024 to do so.

Shifting from traditional security models to the zero-trust model requires significant change. Adoption of zero-trust as a framework requires buy-in at the highest corporate level, significant planning, and deployment strategising (Microsoft, 2023a). Zero-trust adoption and acceptance cannot be defined as a simple pass or fail evaluation, but rather a business and cultural transformation; this must address elements related to governance (architecture and operations), enablers (telemetry, analytics, automation, and orchestration) as well as core domains (identities, workloads, data, networks, and devices) (Mahoutchian et al., 2022).

1.1 Zero-Trust in Corporate Intranets

Implementing the zero-trust security model on the corporate intranet requires a secure means to verify the identity of local servers. On the Internet, trust in the identity of public servers is provided by well-known public Certificate Authorities (CAs). These CAs issue digital certificates that can be used to securely identify a public server (Stapleton and Epstein, 2016). However, it is not possible to naturally use these public CAs to certify servers on the corporate intranet without publicly disclosing sensitive information such as intranet server Domain Name System (DNS) records, as these servers use an internal address space.

To establish trust in the corporate intranet, network administrators are forced to develop or implement their own solutions (often around open-source software), or to rely on third-party solutions (Kerman et al., 2020). Third-party solutions such as endpoint managers can be expensive or labour-intensive to implement, and may lead to unforeseen security vulnerabilities, particularly in multi-node environments (Zeng et al., 2014). Instead, these entities often simply ignore the problem of establishing trust in their networks, thereby leaving servers vulnerable to impersonation, and creating opportunities for information theft and unauthorised access to resources.

2. The Intranet Certification Problem

A problem concerning Zero Trust and *intranet* Hypertext Transfer Protocol Secure (HTTPS) certificates was identified practically by the authors when deploying custom security devices with web interfaces into commercial intranets. Such devices typically do not have public IP addresses, nor public DNS entries. Consequently, such devices cannot naturally present HTTPS certificates that have been signed by a well-known CA without publicly disclosing DNS records that map to internal IP addresses.

The most common approach to mitigating this problem is to use a self-signed certificate (i.e., the device itself signs the certificate). The first time a browser interacts with such a device, the user will be warned that the presented certificate is not publicly signed and must accept this self-signed certificate. This creates an opportunity for a man-in-the-middle (MITM) attack (Conti et al., 2016; De La Hoz et al., 2014). The use of these self-signed certificates also decreases zero-trust metrics (Simpson, 2022).

2.1 Problem Definition

The intranet certification problem is defined as follows: “How can intranet servers offer HTTPS services without introducing additional risks, associated with self-signed certificates, to clients?”.

To qualify as a solution, we require that a candidate ICT system must satisfy all the following requirements:

***R1**: An intranet host can generate a certificate and have it signed by a CA.*

***R2**: A client browser on the intranet will verify and accept this certificate without warning.*

***R3**: The host’s certificate must not be reusable on other intranets.*

***R4**: No private intranet information, such as DNS records, is publicly disclosed.*

***R5**: Client buy-in does not require full control of client endpoints.*

Depending on the requirements of the organisation, complete control of endpoints can be considered a negative or a positive feature. More specifically, the complexity, cost, and privacy issues introduced by endpoint management may be incompatible with the organisation's stance on security.

3. Literature Review

In this section, we consider some well-known ad-hoc candidate solutions to the intranet certification problem, as well as candidate solutions proposed in the literature.

3.1 Ad-Hoc Candidate Solutions

We consider a second ad-hoc candidate solution over and above the candidate solution of using a public CA and disclosing private IP numbers on a public DNS, as discussed in Section 2.

3.1.1 Privately disclose internal IP number via private DNS

One ad-hoc solution is to have the certificate signed by a public CA, but to only publish the DNS record on an internal (private) DNS server. While this solves the problem of public disclosure of private IP numbers and intranet names, this introduces the additional complexity and security risks of hosting a private DNS server. Hosting a private DNS server is a complex undertaking and can increase the risk of DNS attacks, such as DNS tunnelling, spoofing, or cache poisoning (NIST, 2009).

3.2 Candidate Solutions in the Literature

In this section, two relevant candidate solutions from the literature are reviewed.

3.2.1 Extension of SCEP for browsers

The Simple Certificate Enrolment Protocol (SCEP) is an example of a protocol used for issuing and managing digital certificates in networks (Rutishauser et al., 2002). SCEP is typically used in conjunction with a private (internal) CA to automate the certificate enrolment process. In a typical SCEP implementation, an intranet server (typically a device) sends a certificate enrolment request to a SCEP server, which validates the request and issues a digital certificate to the intranet server. The intranet server then uses this certificate as proof of its identity.

SCEP has been instrumental in facilitating automated certificate enrolment, thereby ensuring that certificates are always up-to-date and properly configured (Young and Honore, 2016). However, Rutishauser et al. (2002) identified a limitation of the original SCEP protocol, namely its incompatibility with browsers. This incompatibility is due to unique certificate request formats and the necessity for SCEP clients to access the private key of the certificate request. A consequence of this incompatibility is an inability to automatically install the CA certificate on the browser, thereby not meeting requirement **R2** to solve the intranet certification problem.

To address the above limitation, Rutishauser et al. (2002) proposed extensions to the original SCEP protocol, thus permitting automated CA certificate installation on browsers via a Java-based plug-in. This resulted in an implementation of SCEP client as a Java-based browser plug-in for Netscape Navigator, providing a reference for evaluating the proposed SCEP enhancements. However, these extensions to SCEP were not formally adopted. Furthermore, the proposed extension implementation plug-in was designed for Netscape Navigator only, which was discontinued in 2008.

3.2.2 X.509 home networks

Highlighting the lack of robust security features in networks, Müller et al. (2009) proposed an assisted device registration and service access system for future home networks. This assistance system facilitates the enrolment and distribution of valid X.509 certificates to new devices. These certificates can be used to control network access and ensure secure data transfer, thus addressing some of the key challenges in the implementation of a zero-trust framework in corporate intranets.

The proposed solution requires a home server with a private CA for the home network (Müller et al., 2009). When a new device wishes to register on the home network, software is installed on the new device, which then interacts with the Home Server to sign a generated certificate for the device. This is a manual process in which the Home Server manager interacts with the user of the new device via a side channel to share a PIN. If successful, the Home Server signs the certificate and returns it to the client device. The client then presents the

certificate to use facilities on the home network, such as secure video streaming using Devices Profile for Web Services (DPWS).

This solution does not satisfy requirement **R2** of the intranet certification problem. While the client is authenticated to the services, the services are not authenticated to the client. As such, it does not qualify as a solution to the intranet certification problem.

4. Existing ICT Systems as Candidate Solutions

Three technologies relevant to the intranet certification problem were identified, namely public key infrastructure managers (PKIMs), endpoint managers (EMs), and unified endpoint managers (UEMs). In Sections 4.1 to 4.3, we define these technologies and identify existing ICT systems as examples of each technology.

4.1 PKI Managers

PKIMs facilitate the processes required to implement public key infrastructure (PKI), which involves the use of digital certificates, cryptographic keys, policies, and procedures to facilitate secure communication (OpenXPKI, 2023; Microsoft, 2023b). Tasks such as issuing, revoking, or renewing digital certificates, and managing root CAs or secure key repositories are typically included in PKIM software (DigiCert Inc., 2023). Three certificate enrolment protocols widely used by PKIMs are the SCEP, the Enrolment over Secure Transport (EST) protocol, and the Certificate Management Protocol (CMP).

4.1.1 Microsoft NDES

Microsoft's NDES is a role service within Microsoft Active Directory Certificate Services that offers PKI functionality (Harwood, 2023). NDES acts as a registration authority, enabling the software on routers and other network devices and servers to generate certificates and to have them signed using the SCEP or the CMP. This allows devices without domain credentials to enrol for X.509v3 certificates from an internal CA.

4.1.2 Open-Source PKI managers

The following open-source PKIMs were identified:

- The OpenXPKI Project (OpenXPKI, 2023) is an open-source PKI trust centre. It acts primarily as an online registration authority (RA) or CA to manage X.509v3 certificates. The software provides features that include hosting issuing CAs, utilising SCEP and EST gateways for certificate distribution, hosting a certificate revocation list (CRL), and providing certificate revocation capabilities.
- Dogtag Certificate System (Dewata, 2023) is an open-source CA. It is a collection of technologies that enable the deployment of PKI on a large scale, with features including certificate issuance, revocation and retrieval, CRL generation and publishing, and other functions.
- EJBCA (EJBCA, 2023) is an open-source PKI and CA software that provides functionality such as certificate issuing, management, and validation.

4.2 Endpoint Managers

EMs provide the functionality to monitor and manage multiple nodes or endpoints in a network through automation tools or scripts to increase management efficiency (Zeng et al., 2014). Typically, an EM server is set up within the organisation's network infrastructure. This server hosts software applications that allow administrators to control various aspects of endpoint devices, including software deployment, security configurations, patch management, and remote troubleshooting.

An EM typically installs lightweight agent software on each of the endpoints (Cloudflare Inc., 2023). This software acts as a communication bridge between the endpoint and the management server. It collects data, enforces security policies, and ensures that the endpoint complies with the organisation's IT policies. EMs ensure that only authenticated endpoints can access the network, thus ensuring secure network access.

EMs present functionality to install local root CA certificates on endpoints. However, EMs do not natively offer PKI functionality to generate, issue, revoke, and renew digital certificates, nor manage root CAs.

4.2.1 Microsoft Intune

Microsoft Intune is a cloud-based endpoint management solution (Microsoft, 2023c). It manages user access to organisational resources and simplifies app and device management across multiple devices. These devices include mobile devices, desktop computers, and virtual endpoints, all of which must run Intune client software.

Microsoft Intune can install root CA certificates on endpoints, but requires an existing PKIM to be in place. Specifically, it requires a private (internal) CA and on-premises PKI infrastructure that supports the SCEP or Public Key Cryptography Standards (PKCS).

4.3 Unified Endpoint Managers

Unified endpoint managers (UEMs) combine the functionality of a PKIM (management of a private CA and digital certificates) and an EM (management of endpoints) into a single system (Cloudflare Inc., 2023). These systems can manage server certificates (i.e., certificate issuing, renewal, and revocation), as well as the ability to deploy private root CA certificates onto endpoints and their browsers.

4.3.1 Microsoft Intune paired with Microsoft NDES

Microsoft Intune can naturally integrate with Microsoft NDES to provide full UEM functionality. In this configuration, NDES provides the private (internal) CA, as well as functionality for certificate management. Intune provides endpoint management functionality, as well as the ability to install private root CA certificates on client browsers.

4.3.2 Other UEMs

VMWare Workspace ONE UEM, BlackBerry UEM, and MobileIron UEM are proprietary UEMs that manage and secure endpoints (VMware, 2023a; BlackBerry, 2021; MobileIron, 2023). These devices include mobile devices, desktops, laptops, and Internet of Things (IoT) devices. These UEMs offer certificate enrolment, CA communication, and certificate generation functionality, as well as functionality for installing root certificates on endpoints and their browsers.

5. Results

In this section, the ICT systems identified and described in Section 4 are evaluated with respect to their ability to meet the stated requirements for solving the intranet certification problem, as well as their cost. *Table 1* lists the identified PKIMs, EMs, and UEMs, and indicates whether each candidate solution satisfies each of the requirements **R1** to **R5** of the intranet certification problem. The enrolment protocols used by each system, their pricing, and whether the systems are open source or proprietary are also indicated in the table.

In sections 5.1 to 5.3, we evaluate the shortcomings of PKIMs, EMs, and UEMs generally, with regards to satisfying the requirements of the intranet certification problem. The results are summarised in Section 5.4.

Table 1: Candidate solutions to the intranet certification problem

Name of Solution	Proprietary / Open-Source	Certificate enrolment protocols	Ability to install root certificates on client	PKIM	EM	R1	R2	R3[1]	R4	R5	Pricing (USD)[2]
1. MS Intune (Microsoft, 2023d)	P	SCEP, EST	Y	N	Y	N	Y	Y	Y	N	$4 per user / month (Microsoft Security, 2023)
2. MS NDES[3] (Harwood, 2023)	P	SCEP, EST	N	Y	N	Y	N	N/A	Y[4]	Y	**Perpetual licences:** MS Server Standard: $1069 MS Server Essentials: $501

[1] *If a candidate solution is unable to install root certificates on a client, requirement **R3** will not be applicable as a consequence.*

[2] *Average listed price at time of writing (December 2023).*

[3] *Microsoft NDES is a service offering of Windows Server. Windows Server 2022 has three editions, namely Datacenter, Standard and Essentials. The Datacenter edition is used for highly virtualised datacentres or cloud environments, the Standard edition for physical or minimally virtualised environments, and the Essentials edition for small businesses with up to 25 users and 50 devices (Microsoft Windows Server, 2023a, 2023b).*

[4] *Microsoft NDES can easily be misconfigured to leak intranet DNS information publicly (Heidecker, 2021).*

Name of Solution	Proprietary / Open-Source	Certificate enrolment protocols	Ability to install root certificates on client	PKIM	EM	R1	R2	R3[1]	R4	R5	Pricing (USD)[2]
											Client Access Licences (CALs)[5]: MS Server Standard: $220 per user or device(Microsoft Windows Server, 2023a, 2023b)
3. MS Intune paired with MS NDES	P	SCEP, EST	Y	Y	Y	Y	Y	Y	Y	N	$4 per user / month (Microsoft Security, 2023) + pricing for MS NDES, as indicated above
4. VMware Workspace ONE UEM (VMware, 2023a)	P	SCEP	Y	Y	Y	Y	Y	Y	Y	N	$10 per user / month (VMWare, 2023b)
5. MobileIron UEM (MobileIron, 2023)	P	SCEP	Y	Y	Y	Y	Y	Y	Y	N	$4 per user / month (MobileIron, 2023)
6. BlackBerry UEM (BlackBerry, 2021)	P	SCEP	Y	Y	Y	Y	Y	Y	Y	N	$5 per user / month (BlackBerry UK Limited, 2023)
7. OpenXPKI (OpenXPKI, 2023)	O	SCEP	N	Y	N	Y	N	N/A	Y	Y	Free
8. Dogtag Certificate System (Dewata, 2023)	O	SCEP	N	Y	N	Y	N	N/A	Y	Y	Free
9. EJBCA Community Edition (EJBCA, 2023)	O	SCEP	N	Y	N	Y	N	N/A	Y	Y	Free

5.1 PKI Managers

Candidate solutions 2 and 7 to 9 are PKIMs. Whereas such solutions provide a private (internal) CA and the functionality for servers to generate certificates and have them signed by the private CA, such solutions do not attempt to solve requirement **R2** of the intranet certification problem. This is because the client browser's installed root CA certificates are controlled by the browser or the client operating system itself. It is not the responsibility of the PKI manager to install such root certificates on a client, as PKIMs are server orientated (i.e., not client orientated).

To configure a browser to trust a certificate generated or signed by the PKIM without warning, the user would have to manually add the root certificate of the private CA to the browser's trusted root certificates store. This is a security risk, as it means that the browser will trust the certificate even if it is invalid or has been revoked. Hence, because of the failure to satisfy requirement **R2**, PKIMs cannot be considered a solution to the intranet certification problem on their own.

[5] *A Windows Server Client Access License (CAL) is a license that grants a device or user access rights to services running on a Windows Server operating system* (Microsoft Licensing, 2023).

Candidate solutions 7 to 9 are free, open-source PKIMs. However, technical support is not offered for these PKIMs, and hence their setup and management are complex. In contrast, Microsoft NDES (candidate solution 2) requires Microsoft Server, which incurs a once-off perpetual licence fee of 501 USD (Essentials Edition) or 1069 USD (Standard Edition, for businesses with more than 50 devices or 25 users) (Microsoft Windows Server, 2023b). Microsoft Server Standard Edition incurs a further CAL cost of 220 USD per user or device (Microsoft Windows Server, 2023a).

Due to the popularity and widespread deployment of Microsoft products, we address the particular security risks of candidate solution 2 (Microsoft NDES) below.

5.1.1 Microsoft NDES security risks

Microsoft's NDES infrastructure is designed specifically to provide PKI for public-facing devices (Microsoft, 2023d). In the scenario where devices are used in an *intranet* with private IP addresses and no unique DNS names, multiple issues may arise. Microsoft NDES can allow devices on the Internet or intranet to obtain certificates; this is done by publishing the NDES URL externally from the corporate network. This can be achieved by using different methods, including Azure AD Application Proxy, Microsoft's Web Application Proxy Server, or a third-party reverse proxy. Configuring NDES for Internet or intranet access involves several components and configurations, including Secure Sockets Layer (SSL) certificates, reverse proxies, and URL publication (Heidecker, 2021). This complexity poses challenges for organisations, especially those lacking in-depth expertise, and creates the risk for misconfiguration and the subsequent disclosure of private DNS information.

Furthermore, poor integration with Microsoft NDES can pose significant security risks if it is inadequately hardened or misconfigured. If a malicious actor obtains access to the NDES Enrolment Agent certificate's private key, they could potentially request a certificate valid for logging in to the Active Directory (AD) with any subject. The malicious actor could then log on as a domain administrator using the PKINIT authentication defined in RFC 4556. A compromised AD allows malicious actors to potentially gain access to the network's most vital systems and resources, or obtain administrator privileges, thereby allowing control over the entire domain (Ebad, 2022).

5.2 Endpoint Managers

Candidate solution 1, Microsoft Intune, is exclusively an EM (i.e., it does not provide PKI functionality). In the context of the intranet certification problem, EMs provide a means of installing CA certificates in client browsers. This functionality is essential to satisfy requirement **R2** of the intranet certification problem. However, as EMs mandate full control of endpoints, all EMs fail requirement **R5.** Furthermore, it is not the responsibility of EMs to provide the necessary functionality to issue and manage certificates, which implies that EMs also fail to satisfy requirement **R1**. As a result, EMs cannot be considered a solution to the intranet certification problem.

Microsoft Intune incurs a cost of 4 USD per month for each user of a device (Microsoft Security, 2023). Note that, in order to install the root CA certificate on all devices in the network, the endpoint software needs to be installed on each of these devices.

While some organisations view full control of endpoints as a positive feature, EMs can introduce unexpected security vulnerabilities into an organisation. One example of an unforeseen security vulnerability known to the authors is where intranet web application passwords become device passwords. If the password for the intranet web application is scraped, the entire device is compromised.

5.3 Unified Endpoint Managers

Candidate solutions 3 to 6 are UEMs. Since a UEM is both a PKIM and an EM, it satisfies requirements **R1** to **R4** of the intranet certification problem. However, as an EM, these UEMs fail requirement **R5.** Hence, UEMs do not solve the intranet certification problem.

Microsoft Intune paired with Microsoft NDES (candidate solution 3) will incur the cost of both products, while the remaining UEMs considered in this study (candidate solutions 4 to 6) incur costs of between 5 and 10 USD per user per month; refer to *Table 1*.

5.4 Summary

The study reveals that although the requirements **R1** to **R4** of the intranet certification problem are solved by UEMs, these solutions are expensive. More specifically, UEMs incur a monthly cost per user and / or per client device. In addition, these solutions are also complex to implement and manage. However, as noted, no EM can satisfy requirement **R5** (by definition), and hence no UEM can. Whereas full control of client endpoints may be

compatible with some organisational policies, this potentially introduces additional security vulnerabilities on the clients and privacy concerns.

None of the candidate solutions considered in this study satisfies all the requirements (**R1** to **R5**) of the intranet certification problem. Hence, none of these ICT systems solves the intranet certification problem.

6. Conclusion and Future Work

Solutions to the intranet certification problem that satisfy the relevant technical and security requirements are beyond the reach of smaller, private sector companies and organisations in the domestic public sector, due to their excessive cost and the high level of technical expertise required to implement and manage the relevant systems (Zeng et al., 2014). The analysis of candidate solutions has indicated that EMs require full control of client endpoints, which may be incompatible with company security or privacy policies, in addition to incurring high costs. Whereas PKIMs are available as open-source software, they fail to automatically add the certificate to the browser's trusted root certificates store, thereby requiring manual intervention. This introduces additional security risks and incurs additional management overhead.

UEM solutions such as Microsoft NDES (Microsoft Server) paired with Microsoft Intune can incur costs of over 1000 USD for licencing alone and require a CAL per device or user, costing approximately 200 USD per licence (Microsoft Windows Server, 2023a, 2023b). This conclusion agrees with the authors' experience that smaller private-sector companies and public-sector organisations in underdeveloped and emerging economies often simply ignore the problem of establishing trust in their networks. In doing so, these entities leave their servers vulnerable to impersonation and create opportunities for information theft and unauthorised access to resources (Stankard, 2021).

We conclude that a technological and market gap exists for a secure, cost-effective, and low-complexity solution to the intranet certification problem. A solution that simultaneously meets these criteria promises to accelerate the implementation of the zero-trust model in corporate intranets, thereby contributing towards the safeguarding of these networks. Further, such a solution has the potential to support intranet application and device developers in the ICT domain to implement trust in their networks.

References

BlackBerry, 2021. Securing connections using PKI [WWW Document]. URL https://docs.blackberry.com/content/dam/docs-blackberry-com/release-pdfs/en/blackberry-uem/12_14/administration/Using-PKI-to-secure-connections-BlackBerry-UEM-12.14-en.pdf (accessed 9.4.23).

BlackBerry UK Limited, 2023. BlackBerry Unified Endpoint Management, (UEM) - Industry Leading Security [WWW Document]. URL https://www.applytosupply.digitalmarketplace.service.gov.uk/g-cloud/services/437061688222404 (accessed 12.13.23).

Bobbert, Y., Scheerder, J., 2022. Zero Trust Validation: from Practice to Theory : An empirical research project to improve Zero Trust implementations. In: Proceedings - 2022 IEEE 29th Annual Software Technology Conference, STC 2022. Institute of Electrical and Electronics Engineers Inc., pp. 93–104.

Cloudflare Inc., 2023. What is an endpoint? [WWW Document]. URL https://www.cloudflare.com/learning/security/glossary/what-is-endpoint/ (accessed 7.24.23).

Conti, M., Dragoni, N., Lesyk, V., 2016. A Survey of Man in the Middle Attacks. IEEE Communications Surveys and Tutorials.

De La Hoz, E., Cochrane, G., Moreira-Lemus, J.M., Paez-Reyes, R., Marsa-Maestre, I., Alarcos, B., 2014. Detecting and defeating advanced man-in-the-middle attacks against TLS. In: International Conference on Cyber Conflict, CYCON.

Dewata, E., 2023. Dogtag PKI Documentation [WWW Document]. GitHub. URL https://github.com/dogtagpki/pki/wiki (accessed 9.4.23).

DigiCert Inc., 2023. What is PKI? [WWW Document]. URL https://www.digicert.com/what-is-pki#:~:text=Public%20Key%20Infrastructure%20(PKI)%20is,users%2C%20devices%2C%20or%20services. (accessed 10.20.23).

Ebad, S.A., 2022. Lessons learned from offline assessment of security-critical systems: the case of Microsoft's active directory. International Journal of System Assurance Engineering and Management 13.

EJBCA, 2023. EJBCA CA Concept Guide [WWW Document]. URL https://doc.primekey.com/ejbca/ejbca-operations/ejbca-ca-concept-guide (accessed 9.4.23).

Fortinet, 2023. The State of Zero Trust Report.

Harwood, R., 2023. What is Network Device Enrollment Service for Active Directory Certificate Services [WWW Document]. Microsoft. URL https://learn.microsoft.com/en-us/windows-server/identity/ad-cs/network-device-enrollment-service-overview (accessed 9.4.23).

He, Y., Huang, D., Chen, L., Ni, Y., Ma, X., 2022. A Survey on Zero Trust Architecture: Challenges and Future Trends. Wirel. Commun. Mob. Comput. 2022.

Heidecker, D., 2021. NDES Security Best Practices [WWW Document]. Microsoft Security, Compliance and Identity Core Infrastructure and Security Blog. URL https://techcommunity.microsoft.com/t5/core-infrastructure-and-security/ndes-security-best-practices/ba-p/2832619 (accessed 8.30.23).

Kerman, A., Borchert, O., Rose, S., Division, E., Tan, A., 2020. Implementing a Zero Trust Architecture. NIST Computer Security Resource Center.

Mahoutchian, T., McLaughlin, M., Thayres, A., Rafla, A., 2022. Zero Trust implementation for human centered cyber security [WWW Document]. Deloitte. URL https://www2.deloitte.com/us/en/blog/human-capital-blog/2022/zero-trust-adoption-for-human-centered-cyber-security.html (accessed 6.9.23).

Microsoft, 2023a. Zero Trust adoption framework overview [WWW Document]. URL https://learn.microsoft.com/en-us/security/zero-trust/adopt/zero-trust-adoption-overview#the-zero-trust-adoption-motion (accessed 6.5.23).

Microsoft, 2023b. What is Network Device Enrollment Service for Active Directory Certificate Services [WWW Document]. Microsoft. URL https://learn.microsoft.com/en-us/windows-server/identity/ad-cs/network-device-enrollment-service-overview (accessed 9.4.23).

Microsoft, 2023c. Microsoft Intune securely manages identities, manages apps, and manages devices [WWW Document]. Microsoft. URL https://learn.microsoft.com/en-us/mem/intune/fundamentals/what-is-intune (accessed 9.4.23).

Microsoft, 2023d. Configure infrastructure to support SCEP with Intune [WWW Document]. Microsoft Learn. URL https://learn.microsoft.com/en-us/mem/intune/protect/certificates-scep-configure (accessed 8.30.23).

Microsoft Licensing, 2023. Client Access Licenses and Management Licenses [WWW Document]. URL https://www.microsoft.com/en-us/licensing/product-licensing/client-access-license (accessed 12.13.23).

Microsoft Security, 2023. Discover Microsoft Intune Plans and Pricing [WWW Document]. URL https://www.microsoft.com/en-us/security/business/microsoft-intune-pricing (accessed 12.13.23).

Microsoft Windows Server, 2023a. Windows Server 2022 Remote Desktop Services CAL [WWW Document]. URL https://www.microsoft.com/en-us/d/windows-remote-desktop-server-cal-2022/dg7gmgf0d7hx/0005 (accessed 12.13.23).

Microsoft Windows Server, 2023b. Pricing and licensing for Windows Server 2022 [WWW Document]. URL https://www.microsoft.com/en-za/windows-server/pricing (accessed 12.13.23).

MobileIron, 2023. MobileIron UEM Unified Endpoint Management [WWW Document]. AppDirect. URL https://marketplace.appdirect.com/en-US/apps/263299/mobileiron-uem/editions (accessed 12.13.23).

Müller, A., Kinkelin, H., Ghai, S.K., Carle, G., 2009. An assisted device registration and service access system for future home networks. In: 2009 2nd IFIP Wireless Days, WD 2009.

NIST, 2009. Secure Domain Name System (DNS) Deployment Guide, NIST Special Publication.

OpenXPKI, 2023. Overview [WWW Document]. OpenXPKI Documentation. URL https://openxpki.readthedocs.io/en/latest/reference/configuration/introduction.html (accessed 9.4.23).

Rutishauser, U., Schäfer, A., Müller, A.F., 2002. Open reference implementation of a SCEP v2 client.

Sarkar, S., Choudhary, G., Shandilya, S.K., Hussain, A., Kim, H., 2022. Security of Zero Trust Networks in Cloud Computing: A Comparative Review. Sustainability (Switzerland).

Simpson, W.R., 2022. Zero Trust Philosophy versus Architecture.

Stankard, T., 2021. What is Zero Trust in Terms of Cybersecurity? [WWW Document]. TitanHQ. URL https://www.titanhq.com/blog/zero-trust-cybersecurity/ (accessed 9.4.23).

Stapleton, J., Epstein, W.C., 2016. Security without Obscurity, Security without Obscurity.

VMware, 2023a. Managing Devices with Workspace ONE UEM [WWW Document]. VMWare Docs. URL https://docs.vmware.com/en/VMware-Workspace-ONE-UEM/services/UEM_Managing_Devices/GUID-AWT-MANAGINGDEVICESOVERVIEW.html (accessed 10.18.23).

VMWare, 2023b. VMware Workspace ONE [WWW Document]. URL https://www.vmware.com/products/workspace-one.html#pricing (accessed 12.13.23).

Young, J., Honore, A., 2016. Simple Certificate Enrollment Protocol Overview [WWW Document]. URL https://www.cisco.com/c/en/us/support/docs/security-vpn/public-key-infrastructure-pki/116167-technote-scep-00.html (accessed 12.13.23).

Young, S.D., 2022. M-22-09 Federal Zero Trust Strategy.

Zeng, S., Adam, C., Wu, F., Guo, S., Ruan, Y., Venugopal, C., Puri, R., 2014. Managing risk in multi-node automation of endpoint management. In: IEEE/IFIP NOMS 2014 - IEEE/IFIP Network Operations and Management Symposium: Management in a Software Defined World

Social Media as a Strategic Advantage During Cyberwarfare: A Systematic Literature Review

Errol Baloyi[1], Oyena Mahlasela[1], Nokuthaba Siphambili[1] and Mayan Stegmann[2]
[1]CSIR, Pretoria, South Africa
[2]Nclose, Cape Town, South Africa

EBaloyi2@csir.co.za
OMahlasela@csir.co.za
NSiphambili@csir.co.za
Mayan@Nclose.com

Abstract: In recent years, cyberspace has been shaped by a rapid and transformative technological evolution, which ushered in an era characterised by unparalleled connectivity and innovation. However, this remarkable progress has brought a concerning surge in cyberattacks that have fundamentally altered cyberspace dynamics and refined the nature of contemporary warfare. This refinement was vividly illustrated in the recent Russia-Ukraine conflict, where cyberspace played a pivotal role, blurring the traditional boundaries of conflict in the cyber age. As a result, this study used secondary data to examine how various social media platforms such as Twitter, Facebook, TikTok, and Telegram were used as a strategic advantage during the conflict. The findings disclosed that Russia employed offensive propaganda against Ukraine, while Ukraine adopted a defensive stance, effectively countering the narrative through an active online presence. Moreover, this study underscored the substantial role of social media in warfare and its continued significance in future conflicts. Furthermore, this study provided recommendations for nations to better prepare for such conflicts. The recommendations provide valuable insights to assist decision-makers and policymakers in enhancing cybersecurity awareness and practices within their respective countries.

Keywords: Cyberattacks, Cyberspace, Cybersecurity, Russia-Ukraine, Warfare, Social media

1. Introduction

Since the early 1990s, cyberwarfare has been recognized by its advocates as a groundbreaking military revolution and a highly effective tool in combat, but these was primarily theoretical until the recent Ukraine-Russia conflict where cyberspace was utilized during the conflict (Schulze and Kerttunen, 2023). Even so, several countries have already made notable progress in terms of preparedness, following the Warsaw Summit 2016, where the North Atlantic Treaty Organization (NATO) declared cyberspace as a battlefield, which lead to member countries establishing their respective cyber commands (Smeets, 2023). However, Russia has always demonstrated remarkable proficiency as a cyber power worldwide. Instances of their cyber activities trace back to Moonlight Maze attacks on the United States Department of Defence in the 1990s and suspected involvement in cyberattacks against various targets, including the Estonian government, media, and financial sectors in 2007, Georgian government sites in 2008, Kirgizstan Internet Service Provider attacks in 2009, and Ukrainian government, military, and critical infrastructure attacks in 2014 (Priyono, 2022).

The tensions between Russia and Ukraine have witnessed a significant escalation since 2014, notably characterised by the annexation of Crimea following a contentious local referendum. These tensions have continued to intensify over time, resulting in substantial cyberattacks, particularly targeting Ukrainian power utilities. Moreover, in December 2015 and 2016, over 225,000 individuals in Ukraine experienced power outages due to cyberattacks, with Kyiv also facing similar disruptions (Priyono, 2022). Furthermore, in June 2017, the global impact of the NotPetya cyberattack, attributed to Russia, led to considerable financial losses. After Russia invaded Ukraine on February 24, 2022, cyber warfare played a pivotal role in shaping and influencing international events. This encompassed attacks on communication systems, government websites, border control stations, and Ukraine's digital infrastructure (Przetacznik and Tarpova, 2022). Additionally, Russia employed various tactics, including the manipulation of information through social media, fabricating American military profiles on Twitter to exploit anti-government sentiments and sway populations toward pro-Russian stances (Littell and Starck, 2023; Weinberg, Dawson and Edwards 2023).

2. Methodology

A systematic literature review was conducted to synthesise existing literature on cyberwarfare to understand the use of social media as a strategic advantage during cyberwarfare. The search protocol followed was the Preferred Reporting Items for the Systematic and Meta-analysis (PRISMA) framework (Sarkis-Onofre et al., 2021).

2.1 Identification

Relevant publications were identified through the search terms (*"cyber warfare", "cyberattacks", "cyberspace", "cybersecurity", and "social media").* The search term was then applied to different scholarly databases such as IEEE Digital Library, Scopus, Science Direct and Google Scholar search engine.

2.2 Screening

The publications identified from the scholarly databases were screened, where abstracts were read, and irrelevant studies were eliminated. This was done through the selection criteria applied to get relevant publications—the selection criteria comprised of the inclusion and exclusion criteria as shown in Table 1.

Table 1: Selection criteria exclusion and inclusion

Exclusion	
EC1	Duplicates of the same article retrieved from different database
EC2	Articles that were irrelevant to the study
EC3	Articles that were not written in English
Inclusion	
IC1	Articles that were relevant to cyber warfare and used social media as the strategy
IC2	Articles published between 2022 and 2023

2.3 Eligibility

After applying exclusion and inclusion criteria (n=57), publications were left for eligibility assessment. The eligibility assessment involved reading the whole research article, and the studies found illegible were excluded.

2.4 Included

The publications found legible and considered in the full review are (n=19).

3. Results

Warfare is not new; it has always been a persistent aspect of human history, spanning various conflicts like the Battle of Lepanto, the Napoleonic Wars, the World Wars, The Vietnam War, and so forth (Desai and Manabat, 2022). However, what has evolved is the methods employed in warfare, progressing from primitive wooden spears and rudimentary weaponry to advanced machinery. In the contemporary era, warfare has incorporated a cyber dimension, for instance, using social media to disseminate disinformation during warfare. According to Henkhaus (2022) in the past, disseminating false or misleading content, or even accurate information, required the resources of a state or a powerful actor. However, in the contemporary landscape, anyone can create a TikTok video and instantly share it with a global audience.

3.1 Misinformation and Propaganda

Presently, social media has emerged as the primary platform for disseminating misinformation and conducting propaganda campaigns, often leveraging fabricated content such as false photos, and manipulated videos (Konstankevych et al., 2022). The strategic use of disinformation as a weapon of war seeks to impede international relations, erode public trust in leaders and institutions, and tarnish opposing political ideologies (Morejón-Llamas et al., 2022). Notably, Russia employed this tactic during the conflict. This is because the expansive reach of social media not only enables the instant dissemination of false information but also serves as a powerful tool for influencing public opinion. According to Konstankevych et al. (2022) disinformation model, there are four primary components of disinformation: denial of truth and facts, distortion and twisting of facts, distraction and blame, and Intimidation, causing panic and anxiety. Consequently, this study attempts to align its results to that model to better highlight how social media was used as a strategic advantage during the conflict.

The denial of truth and facts often involves the manipulation of information. Astuti, Attaymini, and Dewi (2022) noted that this manipulation can be characterised as propaganda when used to achieve specific goals. A study conducted by Shultz (2023) is one of the many studies that delve into Russian propaganda, focusing on the top ten topics discussed by Russian government social media accounts. The study found that the primary focus was on the 77th meeting of the United Nations (UN) General Assembly that took place on September 13, 2022. Many tweets from those Russian government accounts engaged in extensive discussions about multilateral relations

and U.N. issues, aiming to depict Russia as a peacekeeping and prominent international player. This communication strategy was recurrently employed by Russian government accounts despite international condemnation of Russia's aggressive actions against Ukraine by the UN.

Following the invasion of Ukraine, Pierri et al. (2023) observed a significant decrease in the prevalence of Russian propaganda. This decline was attributed to the implementation of new policies on various platforms, European regulations targeting Russian propaganda, and Russia's ban on online social networks. However, Górka (2022) cautioned against underestimating Russian influence despite potential setbacks in the information war in Ukraine and the West. In line with this, Geissler et al. (2023) asserted that Russian propaganda has evolved over time, employing new tactics that likely involve an increased use of bots. As defined by Ferrara et al. (2016), social media bots are computer algorithms designed to generate content automatically and engage with humans on social media platforms, aiming to emulate and potentially alter their behaviour.

As social media platforms continue to play an increasingly significant role in people's lives, the influence of bots in shaping individual opinions and public sentiment grows. Particularly when social media becomes a tool for advancing national interests and political propaganda, it can be harnessed to challenge societal and international norms (Shen et al., 2023). Bots offer the capacity to generate many software-controlled social media profiles at a low cost. Even though bots tend to post and receive fewer retweets than human users within social media networks, they still manage to garner more attention than human accounts. Consequently, they can effectively disseminate content that might not gain traction otherwise. A study by Geissler et al. (2023) affirmed the existence of Russian-coordinated campaigns utilising bots, with 20.28% of the information spreaders classified as bots. Many of these bots were created at the outset of the invasion. The study also indicated that bots played an amplifying role in the early dissemination of information. Notably, India, South Africa, and the United States emerged as primary targets for these bots.

Shen et al. (2023) study also supports the findings with a study that examined the popularity of the top 20 hashtags in bot and non-bot tweets, revealing that a significant proportion of social media bots engaged in online conversations during the Russia-Ukraine conflict. Intriguingly, the data showed that tweets from Russian-side bots received a higher average number of retweets, likes, and comments compared to those from the Ukrainian side. Surprisingly, even though Russia was a key participant in the conflict, very few bot accounts identified themselves as being from Russia. Kuźmiński's (2022) further confirmed the presence of Russian propaganda, particularly on video-sharing platforms. Some of the propaganda efforts included the exclusive use of the term "special operation" in Russian narratives, sidestepping accurate descriptions like "war" or "full-scale invasion".

To justify their invasion, the Russian government employed distraction and blame tactics. For instance, they propagated narratives suggesting that Ukraine was a breeding ground for Nazism and had a close relationship with the United States. Additionally, Russia accused Ukraine of stealing gas from "Gazprom" without payment, all to justify and conceal the crimes committed by the Russian military in Ukraine. Russian propaganda sought to distort the course of the war among its citizens and portrayed Ukrainians in a negative light (Konstankevych et al., 2022). Furthermore, the Russian government employed intimidation tactics, causing panic and anxiety among the Ukrainian people. They spread stories suggesting that after the invasion, Ukrainian President Volodymyr Zelenskyy had fled abroad (Astuti et al., 2022). They also alleged that the Ukrainian Armed Forces were sending untrained soldiers into battle and that there was a food shortage in Ukraine (Konstankevych et al., 2022). In a particularly damaging move, a video was circulated that purported to show Ukraine's President surrendering to Russia (Kuźmiński, 2022). Notably, some of these false narratives and videos were shared by verified accounts, adding an additional layer of credibility to the disinformation (Haq et al., 2022).

3.2 Information Dissemination and Mobilizing Public Backing

In the period of transitioning from a looming threat of war to the actual invasion by Russia, Ukrainian leaders, specifically President Zelenskyy, adeptly managed the global demand for information. This skilful handling of information was pivotal to Ukraine's success in information warfare. President Zelenskyy actively engaged in various information campaigns to inform and bolster the morale of the Ukrainian populace. A noteworthy moment occurred in April 2022, approximately six weeks into Russia's invasion of Ukraine, when President Volodymyr Zelensky made a plea at the 2022 Grammy Music Awards, urging the audience to share Ukraine's story on social media, encapsulated by the phrase "Fill the silence" (Serafin, 2022). Subsequent research, exemplified by Sazzed's (2022) study, which involved the analysis of over 80,000 tweets from seven different countries, each representing diverse demographic groups expressing their sentiments and opinions about the conflict, consistently favoured Ukraine, with a prevalent negative sentiment towards the war.

This was a very strategic move by Ukraine, recognising their limitations in cyber capabilities when compared to Russia, prompting them to leverage the power of President Zelensky's speeches and video clips as a formidable means of disseminating information and seeking international assistance. This approach exploited the sympathy and sense of responsibility within the Western world and Ukraine's natural allies (Munk & Ahmad, 2022). Zelensky's omnipresence on various platforms, including TikTok, where "I'm a Zelensky stan" amassed 7.3 million views, and the curiosity surrounding his war attire on Twitter illustrated how he became a source of inspiration. Furthermore, he provided daily updates to the Ukrainian people about the war's status and the necessity of resilience. This exemplified a twenty-first-century form of warfare where information dissemination might outshine physical battlefield actions (Serafin, 2022). However, it's essential to underline that while cyber operations can amplify military efforts, they are most effective when integrated with conventional warfare strategies (Fedotenko, 2023).

Ukraine's Minister for Digital Transformation, Mykhailo Fedorov, also employed a multifaceted strategy to counter Russia's actions. One of his tactics involved reaching out to major tech companies like Meta and Google through Twitter, urging them to block Russian access to their platforms and services. Additionally, Fedorov personally contacted Elon Musk to request assistance in deploying Starlink satellite internet systems to enhance communication capabilities across Ukraine. Furthermore, in the early stages of the invasion, Fedorov took an unprecedented step by issuing a "call to arms" for a civilian cyber army to bolster Ukraine's defence. This call to action was conducted on the encrypted messaging app Telegram. It's worth noting that Telegram served a unique role in this context; while Facebook was limited by legal constraints, Telegram became a hub for activities that fell into the realm of cyberwarfare. Specifically, Telegram groups were utilised to coordinate and organise cyberattacks on Russian websites and services (Ronzhyn, Cardenal, and Batlle, 2023; Serafin, 2022).

According to Ronzhyn et al. (2023), fundraising also emerged as the most prevalent activity on social media during the conflict. Their study analysed Facebook profiles engaged in fundraising and found that while most profiles existed before the onset of the war, they experienced significant growth during the conflict. Similarly, Pohl et al. (2022) delved into analysing 8.7 million original tweets generated by 2.3 million unique user accounts between February 17 and March 3, 2022. Their findings indicated numerous accounts seemingly took advantage of the Ukrainian government's call for cryptocurrency donations. This highlights the dual nature of social media during conflict. While it can be a platform for noble actions like fundraising, there is also a tendency for exploitation. As a result, it is imperative to consider the following recommendations to address these issues.

4. Recommendations

The use of social media platforms to disseminate disinformation, propaganda, and misinformation has the power to shape public perceptions, influence political dynamics, and even impact national security (OECD, 2022). This section delves into recommendations that can empower policymakers, organisations, and individuals to combat the adverse effects of these information warfare campaigns. The reality is that, during future conflicts, it will be crucial to prepare nation-states to employ comprehensive legal frameworks. While respecting free speech, enhancing media literacy, and moderating content in the private sector. Individual users should become vigilant while informed about the conflict. As a result, by providing these recommendations, the authors aim to foster a more informed and resilient society that can better navigate the complex landscape of information warfare in a time of conflict.

4.1 The Role of the Policymaker at a Nation-State Level

During future conflicts, at a nation-state level, policymakers will likely find themselves at the forefront of the battle against disinformation and propaganda. Therefore, this study recommends the development of a framework that will be responsible for addressing the emergence of disinformation and propaganda whilst considering the rights of citizens. The European Parliament briefly highlighted this responsibility by stating that every government has a duty to uphold the Universal Declaration of Human Rights (UDHR), which Article 19 of the UDHR states that: ‘Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers (Colomina, Margalef, and Youngs 2021). Similarly, South Africa and other nations have such regulations. Hence, the goal is to strike a delicate balance by developing and enforcing a new framework that addresses disinformation while upholding fundamental principles of free speech. This framework must provide clear guidelines to distinguish between protected free expression and harmful disinformation, ensuring the latter is appropriately addressed without infringing upon the rights of citizens.

Furthermore, Levush (2019) proposes that policymakers should establish specialised government agencies or units with a mandate to monitor and detect disinformation campaigns. Hook and Verdeja (2022) support this notion by stating that these agencies should be equipped with advanced analytics tools to identify patterns and tactics employed by malicious actors, enabling swift detection and response to emerging disinformation threats. The United Kingdom's (UK) government communication service has already made headway by developing the RESIST 2 Counter Disinformation Toolkit, which is responsible for detecting and responding to online disinformation and fake news campaigns (Pamment, 2021). Therefore, we also recommend that policymakers must also invest in technology and expertise to detect and combat automated bots and troll accounts, which are frequently used to amplify disinformation. Additionally, collaborative and information-sharing partnerships with other nations and international organisations should be prioritized for tackling cross-border disinformation threats, as the interconnected nature of social media necessitates global cooperation and collective responses.

4.2 Countering Disinformation in the Private Sector

According to the UN, the private sector has a non-binding responsibility to respect human rights (UN, 2016). As a result, the largest social media platforms are increasingly implementing policies and functions to protect the human rights of their users and counter the spread of disinformation (Colomina et al., 2021). Concurrently, organisations must invest in media literacy training for their employees, especially those managing social media accounts or interacting with the public (Van Audenhove et al., 2018). Furthermore, it is essential to implement content moderation and reporting mechanisms to promptly remove or label false information on their social media accounts and platforms. Additionally, the private sector organizations should develop a crisis communication plan that outlines protocols for responding to disinformation campaigns targeting their organisation and coordinate response. Furthermore, collaboration with fact-checking organisations, governmental agencies, and other organisations for information sharing and resource pooling is also recommended. Transparency in communication is essential for building trust with the audience.

4.3 The Role of the Individual as an End User of Social Media

According to Colomina et al. (2021), individuals as the end user of social media, play a critical role in countering disinformation online. Therefore, development of media literacy is a necessary first step in being able to identify misinformation, disinformation, and propaganda. As a result, we recommend enrolment of online courses and resources to enhance individual’s media literacy skills. Furthermore, fact-checking is another essential practice, where individuals verify the information before sharing it by using reputable fact-checking websites and avoiding disseminating unverified information (Apuke et al., 2022). Likewise, employing critical thinking, questioning the source of information, and maintaining scepticism toward sensational or biased content are good practices. Moreover, individuals should verify the legitimacy of websites by examining URLs and domains while being cautious of fake news sites that mimic credible news organisations. Thus, reporting false information through social media platform reporting tools is also encouraged. Importantly, individuals should refrain from engaging with, liking, or sharing disinformation and instead report it or inform others about its falsehood. Lastly, cultivating a diverse news feed by following various news sources and opinions helps individuals avoid living in echo chambers and gain a balanced perspective on current events.

5. Discussion

This study exhibited that social media has transcended its role as a mere communication tool and is now emerging as a potent weapon capable of shaping public opinion, influencing perceptions, and ultimately steering the course of events. This transformation was remarkably observed during the recent Russian-Ukraine conflict, where cyberspace played a crucial role from the outset. The incorporation of a cyber element in this conflict serves as a clear indication that the future of conflicts will likely involve cyber strategies, underscoring the necessity for nations to develop a resilient cyber posture. Furthermore, this study highlighted Russia’s offensive stance in the conflict, employing tactics such as disinformation and propaganda attacks against Ukraine to manipulate public opinion. The use of terms like "special operation" instead of "war" aimed at swaying public sentiment in favour of Russia. Despite these efforts, Ukraine successfully maintained a defensive position by maintaining a robust online presence. Leveraging social media, Ukraine also enlisted support from private companies and assembled an army of hackers to counteract Russia's formidable capabilities. Nevertheless, Russia persisted in its efforts. Despite facing restrictions from various companies, they resorted to the use of bots to perpetuate their propaganda and continue shaping public opinion against Ukraine. This underscores the

evolving landscape of warfare, where the role of social media is increasingly significant, necessitating strategic responses from nations to safeguard against such cyber threats.

6. Conclusion and Future Work

The spreading of propaganda and misinformation through social media during warfare represents a new dimension in warfare. Consequently, this study conducted a systematic literature review with a primary focus on the multifaceted role of social media in modern warfare, homing in on its specific utilisation during the Russia-Ukraine conflict. Furthermore, this study aimed to answer a critical research question: how was social media strategically leveraged during this conflict? This was because the Russia-Ukraine conflict served as a stark reminder that warfare extends beyond its conventional realms in the contemporary theatre of war. To comprehensively address the question, this study drew upon a diverse body of literature, synthesising valuable insights to dissect the intricate ways in which social media was and can continue to be employed as a strategic advantage in modern warfare. Beyond exploring its potential benefits, this study delved into the complex consequences of social media in warfare, shedding light on its multifaceted challenges. Furthermore, this study expanded its scope to examine countermeasures that can be employed to mitigate these effects, recognising the importance of a proactive approach in safeguarding national security interests.

The findings generated by this study hold significant implications for governments, law enforcement agencies, and national defence forces, by offering valuable insights to aid in their preparedness and strategic responses to the evolving dynamics of modern warfare. However, it is crucial to acknowledge that this study was constrained to secondary data. The authors recommend that for future work, it is crucial to broaden the scope and include primary data, as this will allow for a more comprehensive understanding of the pivotal role played by social media in the conflict and its ongoing influence on shaping future conflicts.

References

Apuke, O.D., Omar, B. and Tunca, E.A. (2023) "Effect of fake news awareness as an intervention strategy for motivating news verification behaviour among social media users in Nigeria: A quasi-experimental research", *Journal of Asian and African Studies*, Vol 58, No. 6, pp 888-903.

Astuti, Y.D., Attaymini, R. and Dewi, M.S. (2022) "Digital Media and War: Social Media as a propaganda tool for the Russia-Ukraine conflict in the post-truth era", *Proceedings of the Annual International Conference on Social Science and Humanities (AICOSH 2022*), pp 19–27. doi:10.2991/978-2-494069-87-9_4.

Colomina, C., Margalef, H.S., Youngs, R. and Jones, K. (2021) *The impact of disinformation on democratic processes and human rights in the world*, Brussels: European Parliament.

Desai, D. and Manabat, H. (2022) "CyberWarfare: The Past, The Present and the foreseeable future", CPSC - 558 Advanced Computer Networking [Preprint].

Fedotenko, K. (2023) "Cyber warfare as part of information warfare of Russia against Ukraine since the beginning of the 2022 Russian invasion", *Věda a perspektivy,* Vol 8, No. 27, pp 351-357.

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.

Geissler, D., Bär, D., Pröllochs, N. and Feuerriegel, S. (2023)" Russian propaganda on social media during the 2022 invasion of Ukraine", EPJ *Data Science*, Vol *12, No.* 1, pp 35.

Giannaris, P.S., Karamanoli, V., Agathocleous, A., Ilias, I. and Doukas, N. (2022) "Text similarity study for Twitter-based news on Russian-Ukraine cyber war", Paper read at 2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), pp 1-8, Athens, Greece, December.

Górka, M. (2022) "A definitional framework for cyber warfare. The Ukrainian aspect", *Polish Political Science Yearbook*, Vol 51, No .1, pp 1–14.

HHS Cybersecurity Program, (2022). The Russia-Ukraine Cyber Conflict and Potential Threats to the US Health Sector, [online], Administration for Strategic Preparedness and Response (ASPR), https://www.aha.org/system/files/media/file/2022/06/HHS-TLP-White-The-Russia-Ukraine-Cyber-Conflict-and-Potential-THreats-to-the-US-Health-Sector-3-1-22.pdf.

Haq, E.U., Tyson, G., Braud, T. and Hui, P. (2022) "Weaponising social media for information divide and warfare", *In Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, Barcelona, Spain, July, pp 259-262.

Henkhaus, L. (2022) "The role of the internet in Ukraine's Information War", [online], Texas A&M Today, https://today.tamu.edu/2022/03/14/the-role-of-the-internet-in-ukraines-information-war/.

Levush, R. (2019) Government Responses to Disinformation on Social Media Platforms: Argentina, Australia, Canada, China, Denmark, Egypt, European Union, France, Germany, India, Israel, Mexico, Russian Federation, Sweden, United Arab Emirates, United Kingdom. Publication, pp 1–177.

Littell, J. and Starck, N. (2023) "Russian influence operations during the invasion of Ukraine", *International Conference on Cyber Warfare and Security*, Vol 18, No. 1, pp 209–217.

Konstankevych I., Kostusiak N., Shulska N., Stanislav O., Yelova T., and Kauza I. (2022) "Media Manipulation as a Tool of Information Warfare: Typology Signs, Language Markers, Fact Checking Methods", *Ad Alta,* Vol 12, No. 2, Spec. Iss. XXIX, pp 224–230.

Kuźmiński, A. (2022) "The methods of disinformation in the Russia-Ukraine war", *Rhetoric and Communications*, Vol 53, pp 167–171.

Morejón-Llamas, N., Martín-Ramallal, P. and Micaletto-Belda, J.P. (2022) Twitter content curation as an antidote to hybrid warfare during Russia's invasion of Ukraine, *Profesional de la información*, Vol *31, No.* 3.

Munk, T. and Ahmad, J., 2022. "I Need Ammunition, Not a Ride": The Ukrainian Cyber War". *Comunicação e sociedade*, Volu 42, pp 221-241.

OECDb., (2023) Principles of good practice for public communication responses, [online], OECD, https://www.oecd.org/gov/open-government/good-practice-principles-for-public-communication-responses-to-misinformation-and-disinformation.htm

Paavola, J., Helo, T., Jalonen, H., Sartonen, M. and Huhtinen, A.M. (2016) "Understanding the trolling phenomenon: The automated detection of bots and cyborgs in the social media", *Journal of Information Warfare*, Vol 15, No. 4, pp 100-111.

Pamment, J. (2021) "RESIST 2 Counter Disinformation Toolkit", Rep, UK Government Communication Service.

Priyono, U., 2022." Cyber Warfare as Part of Russia and Ukraine Conflict", *Jurnal Diplomasi Pertahanan, Vol* 8, No. 2, pp 44-59.

Pierri, F., Luceri, L., Jindal, N. and Ferrara, E. (2023), April. "Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine", In *Proceedings of the 15th ACM Web Science Conference 2023,* pp 65-74.

Pohl, J., Seiler, M.V., Assenmacher, D. and Grimme, C. (2022" A Twitter Streaming Dataset collected before and after the Onset of the War between Russia and Ukraine in 2022, *Available at SSRN 4066543*.

Przetacznik, J. and Tarpova, S. (2022) "Russia's war on Ukraine: Timeline of cyberattacks", *European Parliamentary Research Service (EPRS)*, pp 1-7.

Ronzhyn, A., Cardenal, A.S. and Batlle, A. (2023) "Collective action on Facebook and telegram during the Russia–Ukraine war", *European Conference on Social Media*, Vol 10, No.1, pp 223–230.

Sazzed, S. (2022) "The Dynamics of Ukraine-Russian Conflict through the Lens of Demographically Diverse Twitter Data", Paper read at 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022, pp 6018-6024.

Sarkis-Onofre, R., Catalá-López, F., Aromataris, E. and Lockwood, C. (2021) "How to properly use the PRISMA Statement". *Systematic Reviews,* Vol 10, No. 1, pp 1-3.

Serafin, T. (2022) "Ukraine's president Zelensky takes the Russia/Ukraine war viral", *Orbis*, Vol 66, No. 4, pp 460–476.

Shen, F., Zhang, E., Zhang, H., Ren, W., Jia, Q., & He, Y. (2023). Examining the differences between human and bot social media accounts: A case study of the Russia-Ukraine War. *First Monday*, *28*(2). https://doi.org/10.5210/fm.v28i2.12777

Schulze, M. and Kerttunen, M. (2023) "*Cyber operations in Russia's war against Ukraine: Uses, limitations, and lessons learned so far",* Berlin: Stiftung Wissenschaft und Politik (SWP).

Shu, K., Bhattacharjee, A., Alatawi, F., Nazer, T.H., Ding, K., Karami, M. and Liu, H. (2020) "Combating disinformation in a social media age", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol *10, No.* 6, pp 1385.

Shultz, B. (2023) "In the spotlight: The Russian government's use of official Twitter accounts to influence discussions about its war in Ukraine", In ACM International Workshop on Multimedia AI against Disinformation (MAD '23), Thessaloniki, Greece, June.

Smeets, M., (2023) "The challenges of military adaptation to the cyber domain: a case study of the Netherlands", *Small Wars & Insurgencies, Vol* 34, No. 7, pp 1343-1362.

Stănescu G. (2022) "Ukraine conflict: the challenge of informational war", *Social Sciences and Education Research Review*, Vol 9, No. 1, pp 146-148.

Talwar, S., Dhir, A., Singh, D., Virk, G.S. and Salo, J. (2020) "Sharing of fake news on social media: Application of the honeycomb framework and the third-person effect hypothesis", *Journal of Retailing and Consumer Services*, Vol 57, pp 102197.

UN., (2022) "Freedom of expression is key to countering disinformation", [online], UNHR, https://www.ohchr.org/en/stories/2022/11/freedom-expression-key-countering-disinformation.

Van Audenhove, L., Marien, I., Craffert, L. and Grove, W. (2018) South Africa's e-Skills Policy, From e-Skills to Media Literacy?, Paper read at *2018 IST-Africa Week Conference (IST-Africa)*, pp 1-11, Gaborone, Botswana, May.

Walker, S., Mercea, D. and Bastos, M. (2021) "Introduction to the disinformation landscape and the lockdown of Social Platforms", *Disinformation and Data Lockdown on Social Platforms*, pp 1–13.

Weinberg, D., Dawson, J. and Edwards, A. (2023) "How the Russian Influence Operation on Twitter Weaponized Military Narratives", *Proceedings of the 18th International Conference on Cyber Warfare and Security*, Vol 18, No. 1, pp 431-439.

The Role Industry-Academia Partnerships Play in Cybersecurity: Exploring Collaborative Approaches to Address Cybercrime

Emmanuel Berkoh, Benjamin Yankson, Mubarak Hussein and Yvonne Dadson
HackIoT Lab – College of Emergency Preparedness, Homeland Security and Cybersecurity, University at Albany, USA

Byankson@albany.edu

Abstract: Cybercrime presents a pervasive threat globally, affecting governments, organizations, and individuals. Addressing this threat requires collaborative efforts, particularly between industry and academia. This paper delves into the key role played by industry-academia partnerships in elevating cybersecurity training and awareness, intending to narrow gaps and align the trajectories of cybersecurity professionals. Employing a secondary research methodology, this study provides insights into the impact of collaborations between academia and industry on Cybersecurity education and awareness. It identifies areas within the education sector that can be improved to enhance cybersecurity awareness. The findings emphasize the crucial role of industry-academia partnerships in advancing cybersecurity awareness and resilience, offering potential solutions for cultivating skilled cybersecurity professionals. Additionally, the research aims to contribute to policymaking by advocating for laws and regulations that encourage collaborations between state institutions and industry to mitigate cybersecurity crime effectively.

Keywords: Cybersecurity, Cybercrime, Training, Collaboration

1. Introduction

The rapid evolution of technology has brought forth significant benefits to society but also heightened vulnerabilities, particularly concerning Cybercrime (Jang-Jaccard & Nepal, 2014). Cybercrime poses substantial threats to various sectors, impacting economic, financial, and information domains and endangering national security interests (Olmstead & Smith, 2017). Consequently, strategic partnerships between academia, government agencies, and industry are imperative to introduce cybersecurity programs and curricula across education sectors, notably higher education, fostering increased awareness, training, and the production of cybersecurity professionals capable of addressing emerging threats (Barati & Yankson, 2022). As documented by Statista, internet-based crimes, including Cybercrime, are among the fastest-growing cybersecurity threats, underscoring the urgent need for comprehensive collaboration and educational initiatives (Perosyan, 2023).

The financial ramifications of Cybercrime are staggering, with organizations worldwide suffering significant economic losses and data breaches amounting to billions of dollars (Barati & Yankson, 2022). Data breach costs vary across industries, with the healthcare sector bearing the highest average cost per breach, highlighting the criticality of industry-academia partnerships in mitigating such risks (Petrosyan & Ten, 2023). The inherent challenges arising from the rapid adoption of technology necessitate collaborative frameworks between industry, government, and academia to develop long-term solutions that minimize associated risks (Brooks, 2023).

The complexity of modern information systems, compounded by factors like cloud computing and edge computing, underscores the need for enhanced collaboration to address evolving threats effectively (Brooks, 2023). Addressing pressing Cybersecurity issues requires a cohesive approach that leverages the expertise and resources of all stakeholders, emphasizing shared responsibility and coordinated efforts (Brooks, 2023). Collaborative cybersecurity efforts should not be confined to research, government, or industry alone but should encompass a holistic, collaborative approach to safeguarding technology-dependent environments (Brooks, 2023). As Cybercrime's impact continues to escalate, collaboration becomes increasingly critical, with global costs projected to reach trillions of dollars annually by 2025 (Brooks, 2023). Figure 1 and Figure 2 illustrate the yearly costs of data breach trends and some of the World's most significant data breaches, respectively. For instance, McCandless & Evans study reveals the magnitude and scope of cybersecurity challenges (McCandless & Evans, 2022).

Table 1: The Healthcare industry-associated average data breach cost (Alda, 2023)

Categories	May 2020 - Mar 2021	Mar 2021-Mar 2022	Mar 2022-Mar 2023
Healthcare	9.23	10.1	10.93
Financial	5.72	5.97	5.9
Pharmaceuticals	5.04	5.01	4.82
Technology	4.88	4.97	4.66
Energy	4.65	4.72	4.78
Professional Services	4.65	4.7	4.47
Industrial	4.24	4.47	4.73
Global Average	4.24	4.35	4.45
Research	3.6	3.88	3.63
Education	3.79	3.86	3.65
Consumer	3.7	3.86	3.8
Entertainment	3.8	3.83	3.62
Communications	3.62	3.62	3.9

Data analysis attributed to Statista data from March 2022 and March 2023 indicates the highest average data breach cost in Healthcare is approximately 11 million USD. The average per breach in the financial sector was 5.9 million USD. Table 1 provides data breach costs for Pharmaceuticals, Technology, Energy, Professional Services, Industrial, Global, Research, Education, Consumer, Entertainment, and Communications (Petrosyan & Ten, 2023).

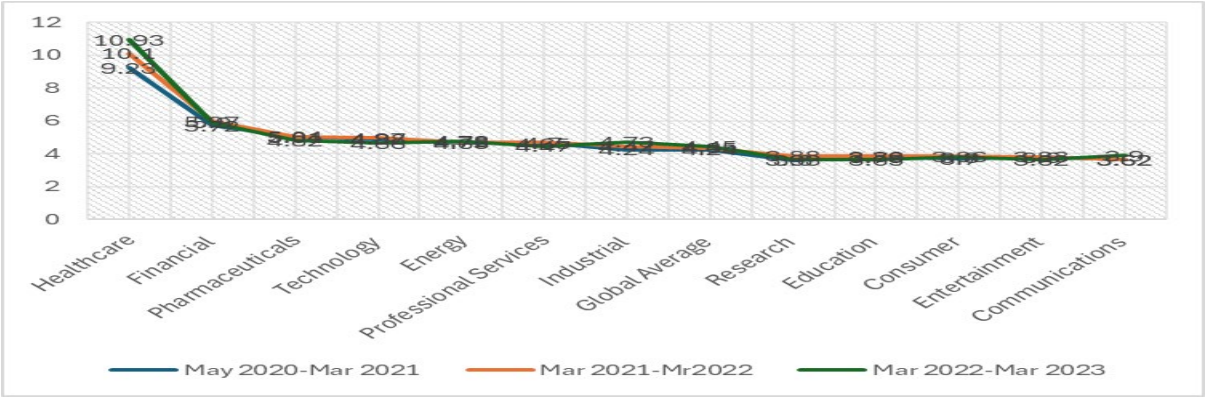


Figure 1: Yearly costs of data breach trends and some of the World's most significant data breaches

The above graph depicts a yearly trend in costs of data breaches within specific industries relative to the global average costs.

1.1 Challenges

The introduction, integration, and rapid adoption of technology spanning innovation in multiple fields have demonstrated significant benefits, including efficiency, productivity, and flexibility with society (Brooks, 2023). Such technological benefits are also accompanied by the risk of harm, especially to the system that supports our critical day-to-day activities, which can be subject to disastrous results such as financial loss, reputation loss, and, in some cases, significant safety concerns as we become more technologically dependent. For example, the adoption of ubiquitous Internet of Things devices, the adoption of AI-enabled devices, the integration of 5G, and the adoption of machine learning are resulting in paradigms requiring these systems to process vast amounts of data or system dependency and requiring instrumentation to protect these dependent systems from breach. Such a problem requires a solution that considers active and current research, industry expertise, and government working together to develop meaningful long-term solutions to minimize the risk associated with integrating and using new technology (Brooks, 2023).

The demand for new technology has become a competitive advantage for companies and a nation-state, which, in some cases, is mostly transitioning rapidly with increased vulnerability and related risks to cause significant

risk. Fundamentally, the continued adoption and use of technology in all aspects of everyday life have increased the attack, which has been enlarged as many companies integrate complex use of systems that allow interconnectivity and people to work anywhere across the globe. Such has resulted in a very complex IT permitter compared to traditional on-premises solutions. For example, today's information systems design and solutions, such as cloud and edge computing, need better industry and academic collaboration to design solutions that address threats through detection, analysis, and response.

To address some of the key pressing Cybersecurity issues in adopting new technologies, success can be based on an articulated approach that can be based on proper collaboration between industry, government, and academia-based framework that depends on the proper articulation of Cybersecurity problems, data availability, and sharing, investment in design and protecting new and emerging technologies). Such framework requires the development and facilitation of by industry, government, and academic entities with properly defined roles that develop continued working partnerships with the sole goal of addressing risk to information systems and avenues to leveraging resources and expertise in the respective sector to mitigate risk to widely used and adapted information system for everyday use or critical sectors. Expertise and resource coordination in addressing Cybersecurity challenges based on sharing resources such as tools, methods, and research results will be essential in such collaboration. Brooks (2023) states cybersecurity problems should not be isolated to research, government, or industry but rather collaborative effort. Our era fully depends on technology, which can be subject to Cybercrime and impact the brand, reputation, and revenue significantly. Such impact is well documented through continued attacks and costs related to such attacks. For example, as per estimates, it will grow by 15 percent per year, reaching USD 10.5 trillion annually by 2025 and costing the global economy over six trillion dollars (Brooks, 2023).

Over the last few years, high-profile cybercrime cases that attracted the attention and imagination across the globe include the hacking of the World Anti-Doping Agency and the subsequent release of medical testing records for several athletes and the hacking of Yahoo servers which led to access to confidential information of more than 1 billion account holders. In 2017, the United States Department of Homeland Security, FBI, and CIA all testified and stated that the Russian government was involved in a scheme to hack and influence the outcome of the 2016 US presidential elections (Olmstead & Smith, 2017). These examples demonstrate that the cost of Cybercrime is a social, political, and economic problem that grows daily. As such, a new approach, which entails pursuing strategic industry-academia partnerships, is required to allow the nation to build capacity and resilience to counter the threat of Cybercrime. Ending the ever-evolving threat of Cybercrime through strategic industry-academia partnerships is crucial because it will help bolster national security by preventing potential economic, financial, and personal information loss for millions of people and organizations.

Addressing such cybersecurity-related threats and vulnerabilities and combating or addressing them is necessary to improve the nation's cybersecurity awareness and resilience levels by pursuing robust academia-industry partnerships, bridging the gap, and putting Cybersecurity security researchers and professionals on the same trajectory. This work report will primarily rely on the literature review methodology to obtain information from scholarly articles to help explain the critical role that industry-academia partnerships play in increasing cybersecurity training and awareness. The scope of this study entails exploring the role of industry-academia strategic partnerships in cybersecurity training and awareness. The research investigates the critical role of the education sector in creating awareness around the vital issue of Cybersecurity in a world where Cybercrime is quickly transcending borders and adversely impacting digital systems and devices.

The objective of this research is to (1) examine the role of industry-academia partnerships in cybersecurity training and awareness; (2) To identify the people or organizations adversely impacted by Cybercrime and propose potential academia-industry solutions to help end it; (3) To explain what industry-academia partnerships ought to do to reduce or eliminate the threat of Cybercrime; (4)To determine the critical areas in the education sector, significantly higher education, that require Improvement to build and bolster cybersecurity awareness and resilience. The key research questions include: (1) To what degree do industry-academia partnerships and initiatives influence cybersecurity training and awareness levels? (2) What are the critical areas in the education sector require Improvement to sustain efforts to create a nation embedded in cybersecurity awareness? (3) To what degree do industry-academia partnerships affect or influence the ability to achieve the desired levels of cybersecurity awareness? (4) How will academia-industry partnerships help develop cybersecurity professionals to bolster the nation's cybersecurity capacity and resilience?

This work contributes to the current literature on industry and academic collaboration in Cybersecurity. Second, this work will serve as an academic resource for other researchers interested in expanding on industry-

academia partnerships and initiatives that influence cybersecurity training and awareness levels by addressing critical areas in academia that require Improvement to sustain efforts to create a nation embedded in cybersecurity awareness. The rest of the paper is organized as follows: Section 2.0 Background and related work. Section 3.0 presents a detailed methodology and implementation. Section 4.0 provides findings and analysis, section 5.0 concludes, and section 6.0 contains the references.

2. Background and Related Work

Globally, Cybercrime poses an urgent and growing threat to governments, businesses, and individuals. The global cost of Cybercrime in 2021 was estimated to be over $6 trillion, up from around $3 trillion in 2015 (Morgan, 2018, 2020). In 2023, the amount exceeded $8.15 trillion; more than $13.82 trillion is estimated in cybercrime costs by 2028 (Petrosyan, 2023). Increasingly, cybercriminals use sophisticated methods to breach systems and steal valuable data and funds. As the digital landscape expands with new technologies like cloud computing, IoT, and cryptocurrencies, so do vulnerabilities and opportunities for Cybercrime (Phipps, 2024). However, no single entity can address the existing complex and evolving challenges of Cybersecurity alone. Therefore, collaboration between key stakeholders, including industry, government, and academia, is critical for developing holistic and proactive solutions (The White House, 2023). In particular, industry-academia partnerships hold great promise for bridging gaps between research and practice to enhance Cybersecurity (Ahmed et al., 2022).

A recent review of the literature on industry-academia collaborations for Cybersecurity reveals several benefits and focus areas for these partnerships. Partnerships can help advance cybersecurity research and translate findings into impactful real-world solutions by bridging gaps between academic research and industry needs (Smyth et al., 2019). While universities often lack the practical context, companies frequently lack resources for foundational research. Activities like collaborative projects, staff exchanges, data sharing, and joint training programs can overcome these challenges (Ahmed et al., 2022). Strategic policies can catalyze impactful collaborations between industry and academia, but a measured approach may be required (World Economic Forum, 2020).

Vogel (2016) writes that industry-academia partnerships have a significant role in improving knowledge and awareness around the emerging issue of Cybersecurity. Strategic collaborations between government, academia, industry, and the private sector can breach the cybersecurity skills gap by ensuring that relevant cybersecurity skills are taught to students across all levels. The private sector and academia can collaborate to establish a US Cybersecurity challenge aimed at augmenting the pool of IT and cybersecurity professionals in the United States. This initiative seeks to bolster the nation's capacity to address and mitigate emerging cyber threats through proactive measures. Additionally, the author advocates for implementing cybersecurity challenges, competitions, and networking events to broaden access to career pathways within the cybersecurity sector. Such endeavors are pivotal in bridging the gap identified in the author's work. Pengulu, Lee & Muller (2012) posit that strategic partnerships between federal, state, and local governments, industry organizations, territorial-level private companies, and academic institutions can address cybersecurity issues.

Cooperation between the private sector and academia is a highly effective tool for achieving cybersecurity goals. The federal government has pursued collaborative efforts with academia and the private sector, bolstering the nation's cyber defense capacities. Encouraging the growth of these collaborations will help combat and eliminate the threat of Cybercrime for the years to come. Members of academia and the private sector should also be involved in the efforts to protect the nation's critical digital infrastructure by sharing information that can help prevent hostile acts of Cybercrime with them. In the past, InfraGard, an FBI program, has collaborated with the private sector and academia to increase the level of awareness on matters of Cybercrime and counterterrorism. Expanding such partnerships to involve the academic sector can improve the nation and make it more proactive in facing the threat of Cybercrime. Although the authors identify critical issues in such partnerships, they missed an opportunity to identify them.

Plunkett (2014) also affirms that robust partnering between the National Security Agency, government agencies, and academia can help build and develop the cyber workforce. Such partnerships provide opportunities for young people to acquire the skills and knowledge required to perform and execute a whole spectrum of cybersecurity and information assurance functions. Plunkett also notes that higher education in Cybersecurity and information assurance must be promoted to enable the nation to produce enough professionals with skills and expertise to address cybercrime vulnerabilities. It is also imperative that the United States and the rest of the World focus and direct their resources toward creating a broader and more technically diverse human workforce to meet the ever-evolving threat of cybersecurity challenges. The author

also ascertains that the role of academia in achieving cybersecurity awareness is critical. In partnerships with the NSA and the private sector, academia can help build strong cybersecurity curriculums that can benefit the nation from a cybersecurity perspective. The academic sector can also help develop an acceptable and desired level of cybersecurity resilience by embedding cybersecurity programs and curriculums in higher education.

Wang & Guo (2020) postulate that industry-academia partnerships- collaborations between the Department of Homeland Security, the Department of Education, the Department of Defense, the Office of the Director of National Intelligence, and the National Science Foundation- can help increase the nation's awareness of cybersecurity threats. Introducing cybersecurity education programs and higher education curricula can help train and produce cybersecurity professionals, cybersecurity researchers, and a country full of citizens aware of Cybersecurity. One way the education sector can improve the nation's Cybersecurity professional capacity is by holding annual cybersecurity championships that bring together all the institutions of higher learning. Such encourages college and university students to actively participate in cybersecurity construction and produce enough cybersecurity specialists to meet the growing needs of the information defense system. More importantly, countries worldwide should strive to provide excellent and sustainable continuing education and training opportunities for cybersecurity educators to improve their skills and knowledge. It is critical, considering that the field of Cybersecurity is evolving rapidly. Thus, according to cybersecurity trainers and educators, a continuous learning opportunity helps update their knowledge reserves, making them competent and qualified cybersecurity experts.

Critical areas to focus partnerships include secure software development, cloud security, network defense, cryptography, and AI/ML security. Such focus helps align research with current industry problems and provides innovative capabilities to improve cyber defenses (Tariq et al., 2023). Industry-academia partnerships also facilitate the development of Cybersecurity through shaping curricula and providing internships and training. The partnership is crucial to addressing the skills gap as companies struggle to find professionals with relevant expertise (Blažič, 2021). According to a 2022 Workforce Cybersecurity study, there were over 700,000 unfilled positions in the United States as companies struggled to recruit professionals with relevant hands-on skills while students lacked an understanding of real-world workforce needs (Eckert, 2023). Furthermore, governments have crucial roles in incentivizing partnerships through funding programs, grants, tax credits, and consortiums. However, some argue that balanced oversight is needed to avoid potential pitfalls, ensure academic independence, and develop sustainable skills programs (Pinguelo & Muller, 2012; Plunkett 2014; Vogel, 2016; Wang & Guo, 2020).

3. Methodology

This research examines the impact of industry-academia partnerships on cybersecurity training and awareness, focusing on collaborative approaches to enhance global cybersecurity measures. The study uses secondary research methodology to explore the roles of industry-academia collaborations in raising awareness and fortifying the World against cyber threats. The primary data collection method involves reviewing and analyzing existing literature from digital library databases, including IEEE Xplore, Google Scholar, and JSTOR. The study employs keywords such as Cybersecurity industry, cybersecurity collaborative approaches, academic partnership, Cybersecurity attacks, Cybercrime, and Cybersecurity research during a comprehensive literature search to identify significant challenges.

The selection criteria for research works are narrowed to those specifically related to collaborative approaches to Cybersecurity and Cybercrime, focusing on papers published between 2018 and 2021. However, the study also incorporates relevant seminal papers predating 2018. Literature sources are selected based on publication year and information quality. The reason for employing a secondary research methodology is multifaceted. Firstly, this approach facilitates evaluating, reviewing, and selecting pertinent literature sources, given their availability and accessibility online. Secondly, the secondary research methodology proves cost-effective, eliminating the need for fieldwork. Access to various academic search engines via an internet-enabled laptop provides the researcher with a wealth of credible and relevant sources.

4. Findings and Discussion

The findings presented from the literature review indicate that academia-industry partnerships play a vital role in raising cybersecurity awareness by empowering learners with skills and knowledge to mitigate the rising cybersecurity vulnerabilities. For example, Vogel (2016) notes that the US government, through the American National Security Agency (NSA), can work with academia to develop cybersecurity programs tailored to equipping young learners with skills and capabilities to enable them to combat the rising levels of Cybercrime.

These programs allow the country to build resilience and capacity in the digital arena. In the past, similar cybersecurity programs rolled out by the National Security Agency (NSA) in partnership with academia have borne the desired outcomes. These programs have also helped promote cybersecurity knowledge and awareness in higher education by producing qualified Cyberdefense professionals. Similarly, Wang & Guo (2020) agree that the increased production of cybersecurity professionals made possible by establishing a sustainable education system that provides education and training opportunities to cybersecurity educators can help reduce the vulnerabilities experienced across the US networks. This is by giving the learners and cybersecurity educators a sustainable and continuing framework that enables them to constantly hone, shape, and improve their cyberspace knowledge and skills.

Notwithstanding the current gaps in industry and academic partnership, there is a growing concern within Cybersecurity as a field, which is a worldwide shortage of qualified personnel who take roles in academia, industry, and government, therefore adding to the challenging risk issue already faced by the field. A recent study by the government and public sector of approximately 3.4 million shortfalls workers demonstrated an increase in shortage of about 26% from 2021 (ISC2, 2022). Although academia continues to train future cyber workers who can problem-solve and understand the design and use of technology and related trends and security risks they face, there continues to be a gap in the number of professionals entering the field (Vogel, 2016). The need for growth in such collaboration in industry, academia, and government is poignant here. The growth in Cybercrime demands workers who can work on cutting-edge technology to protect their respective countries against threats and address vulnerabilities in Information systems against future Cyber-attacks. Growth can be achieved through collaboration between these avenues to provide opportunities for minority women and retooling and skilling opportunities for other professionals, such as veterans.

The current growth of technology and digital resources in all aspects of society pushes the next frontier of vulnerability and threat landscape and avenues for threat actors with malicious intent. Unfortunately, the growth in technology and digital resources that have not yet been adopted into mainstream use or yet to be by the cybersecurity industry, academia, or government poses great and dire challenges and require cooperation and collaboration to begin to unearth vulnerabilities and threat in other to protect organizations, institution, or government. To address the challenge at hand, each of the identified areas has a role to play. Academic institution roles include training and adopting pedagogy to increase the required skills yet maintain the rigor necessary for Cyber professionals to solve complex problems in the field. At the same time, academic institutions must maintain the balance of improving enrollment and recruitment from some nontraditional student cohorts, such as programs and opportunities for minorities. The role also involves contributing to research and development of breakthrough technology to address some of the unresolved complex problems. Although some institutions have continued development in these areas, others have lagged in the research and development sphere. As stipulated by (Brooks, 2023), a school such as MIT, Cal Tech, University of Chicago, Harvard, Carnegie Mellon, and other research-intensive institutions continue to find avenues to conduct basic research and create a pipeline of scientists, engineers, and cybersecurity practitioners who care capable of commercializing the technology.

Resolving this issue is not only isolated to institutions but also the government has a role to play, including but not limited to policy foundation for adoption and use of new technology, promoting public and private partnership, developing best practices, and designing models to prototype and providing working avenues for both academia and industry to contribute to event programs. The government's role also serves as the foundation. It provides some level of resources and funding necessary to jumpstart addressing the pressing issues of vulnerability and threat in current and emerging technologies while at the same time developing comprehensive research and development programs and supporting git.

In the cybersecurity industry, roles include conducting research and development from some of the basic research academia has produced, offering a playback for innovation through some lessons and experiences, and sharing them with the industry for solutions. Industry investment and collaboration are also necessary for developing products and bringing products into being while developing talents. Such industry contributions include technology development, research and development in many areas, and collaboration with academia and government to discuss operational challenges.

Similarly, Blažič (2022) agrees that the education sector can help increase cybersecurity awareness by imparting learners at all levels the desired specialized skills and knowledge to help combat the rising threat of Cybercrime. The author argues that the United States should increase its cybersecurity training and education efforts by adopting new approaches. Strategic partnerships between academia and the private sector are crucial for

promoting and enhancing cybersecurity competence and resilience levels. The ultimate objective of such collaborations is to cultivate a skilled workforce comprising cybersecurity professionals and ICT experts. Blažič's research findings underscore that a significant factor hindering the nation's ability to achieve desired levels of cybersecurity awareness and resilience is the prevalence of theory-based education curricula rather than hands-on training approaches. Consequently, reorganizing and redefining educational and training pathways are essential steps toward bolstering the nation's capacity to cultivate cybersecurity professionals who can effectively address the evolving demands of cyberspace.

Plunkett's findings also affirm that the most effective way of achieving cyberspace's desired confidence and resilience is through robust partnering between academia, the private sector, government agencies, and industry. These partnerships help unveil the much-needed resources and experts that can help shape and mentor young learners, thus enhancing the nation's capacity to build a pool of sustainable cybersecurity professionals capable of meeting the needs and demands of the cyberspace sector. Plunkett (2014) also reveals that such partnerships are critical for creating the next generation of cybersecurity practices and standards, which will play a vital role in educating the future cyberspace workforce. Similarly, the Pinguelo Lee & Muller findings assert that cooperation between academia, the government, and the private industry can significantly help achieve cybersecurity training and awareness objectives and goals. Leaving the burden of achieving the desired levels of cybersecurity awareness and resilience to the education sector alone or the government can be costly. As such, the partnerships will aid in the efforts to help the nation bolster its cybersecurity defenses and capabilities.

5. Conclusion

This study significantly advances cybersecurity awareness and resilience by proposing innovative strategies fostering collaboration among industry, academia, and government entities. The study underscores the pivotal role of academia, in conjunction with government agencies and the private sector, in elevating national cybersecurity awareness. By actively participating in initiatives aimed at producing highly skilled IT professionals with requisite cybersecurity expertise, academia becomes a key driver in achieving this goal. The ongoing technological evolution underscores the escalating reliance on technology across various sectors. Organizations are increasingly adopting advanced technological solutions to harness the benefits of emerging technologies. Coordinating efforts for cybersecurity solutions becomes imperative to safeguard these technologies, thereby enhancing organizational efficiency, profitability, and competitive advantage.

Furthermore, this research highlights the feasibility of overcoming cybersecurity competency gaps, awareness deficits, and skill shortages by reshaping the current educational paradigm. Shifting towards a more practical, hands-on training approach rather than a theory-centric education model proves instrumental in bridging these gaps. Establishing robust strategic partnerships among academia, the private sector, and government agencies emerges as a vital catalyst, providing essential human capital and financial resources. These collaborations are integral to developing and implementing a revised educational curriculum tailored to producing a greater number of cybersecurity experts and professionals, effectively addressing the nation's pressing needs in the realm of Cybersecurity.

References

Ahmed, F., Fattani, M. T., Ali, S. R., & Enam, R. N. (2022). Strengthening the Bridge Between Academic and the Industry Through the Academia-Industry Collaboration Plan Design Model. *Frontiers in Psychology*, *13*. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2022.875940

Alda,M.(2021).TheStatisticsPortal.Statista.https://www.statista.com/markets/424/topic/1065/cyber-crime-security/#definition

Blažič, B. J. (2022). Changing the Landscape of Cybersecurity Education in the EU: Will the
New Approach Produce the Required Cybersecurity Skills? Education and information technologies, 27(3), 3011-3036.

Brooks, C. (2023, October 5). Academia, industry, and government can create innovative partnerships and Help Secure Our Digital Future. Forbes. https://www.forbes.com/sites/chuckbrooks/2023/07/13/academia-industry-and-government-can-create-innovative-partnerships-and-help-secure-our-digital-future/?sh=362c83ca6730

Blažič, B. J. (2021). The cybersecurity labor shortage in Europe: Moving to a new concept for education and training. *Technology in Society*, *67*, 101769. https://doi.org/10.1016/j.techsoc.2021.101769

Eckert, C. (2023). *U.S. Desperately Needs Cyber Talent, Congress Says*. https://www.nationaldefensemagazine.org/articles/2023/6/26/us-desperately-needs-cyber-talent-congress-says

ISC2, I. (2021, February 12). A critical need for cybersecurity professionals persists in the year 2022. https://media.isc2.org/. https://media.isc2.org/-/media/Project/ISC2/Main/Media/documents/research/ISC2-Cybersecurity-Workforce-Study-2022.pdf

Jang-Jaccard, J., & Nepal, S. (2014). A survey of emerging threats in Cybersecurity. *Journal of Computer and System Sciences*, *80*(5), 973–993. https://doi.org/10.1016/j.jcss.2014.02.005

Morgan, S. (2018, December 8). Cybercrime To Cost The World $10.5 Trillion Annually By 2025. *Cybercrime Magazine*. https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

Morgan, S. (2020, November 9). Cybercrime Damages $6 Trillion by 2021. *Cybercrime Magazine*. https://cybersecurityventures.com/annual-cybercrime-report-2017/

Olmstead, K., & Smith, A. (2017, 01 26). Americans and Cybersecurity. Retrieved from Pew

Research Center: https://www.pewresearch.org/internet/2017/01/26/americans-and-cybersecurity/

Pinguelo, F. M., Lee, W., & Muller, B. W. (2012). Virtual Crimes, Real Damages Part II: What Businesses Can Do Today to Protect Themselves from Cybercrime, and What Public-Private Partnerships are Attempting to Achieve for the Nation of Tomorrow. Va. JL & Tech., 17, 75.

Plunkett, D. A. (2014). Achieving Confidence in Cyberspace in an Ever-Changing

Ecosystem. Journal of Information Warfare, 13(2), 1-7.

Petrosyan, A. (2023, December 19). Topic: U.S. consumers and Cyber Crime. Statista. https://www.statista.com/topics/2588/us-consumers-and-cyber-crime/#topicOverview Ani Petrosyan,

Petrosyan, A., & Ten, O. (2023, October 10). The global average cost of a data breach by industry 2023. Statista. https://www.statista.com/statistics/387861/cost-data-breach-by-industry/

Petrosyan, A. (2023). *Global Cybercrime is estimated to cost 2028*. Statista. https://www.statista.com/forecasts/1280009/cost-cybercrime-worldwide

Phipps, B. (2024, January 2). *The Evolution of Cybersecurity: Staying Ahead of Emerging Threats - Bakersfield, Lancaster, Porterville*. Second Star Technologies. https://www.secondstartechnologies.com/2024/01/the-evolution-of-cybersecurity-staying-ahead-of-emerging-threats/

Pinguelo, F. M., & Muller, B. W. (2012). *Virtual Crimes, Real Damages Part II: What Businesses Can Do Today to Protect Themselves from Cybercrime, and What Public-Private Partnerships are Attempting to Achieve for the Nation of Tomorrow* (SSRN Scholarly Paper 2028457). https://papers.ssrn.com/abstract=2028457

Plunkett, D. A. (2014). Achieving Confidence in Cyberspace in an Ever-Changing

Ecosystem. Journal of Information Warfare, 13(2), 1-7.

Smyth, S. J., Curran, K., & McKelvey, N. (2019). The Role of Education and Awareness in Tackling Insider Threats: In I. Vasileiou & S. Furnell (Eds.), *Advances in Information Security, Privacy, and Ethics* (pp. 33–52). IGI Global. https://doi.org/10.4018/978-1-5225-7847-5.ch003

Tariq, U., Ahmed, I., Bashir, A. K., & Shaukat, K. (2023). A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review. *Sensors*, *23*(8), 4117. https://doi.org/10.3390/s23084117

The White House. (2023). *National-Cybersecurity-Strategy-2023.pdf*. https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf

Vogel, R. (2016). Closing the Cybersecurity Skills Gap. Salus journal, 4(2), 32-46.

Wang, W., & Guo, L. (2020, December). Cyber Security Training in Europe and America and its Enlightenment to China. In 2020, the 6th International Conference on Social Science and Higher Education (ICSSHE 2020) (pp. 377-382). Atlantis Press.

Wang, W., & Guo, L. (2020). Cyber Security Training in Europe and America and its Enlightenment to China: *Proceedings of the 2020 6th International Conference on Social Science and Higher Education (ICSSHE 2020)*. 2020 6th International Conference on Social Science and Higher Education (ICSSHE 2020), Xiamen, P.R. China. https://doi.org/10.2991/assehr.k.201214.073

World Economic Forum. (2020). *Partnership Against Cybercrime*. World Economic Forum. https://www.weforum.org/projects/partnership-against-cybercime/

World Economic Forum. (2022, December 2). *How to Develop the Global Cybersecurity Workforce Today*. World Economic Forum. https://www.weforum.org/agenda/2022/12/how-to-develop-the-global-cybersecurity-workforce-and-build-a-security-first-mindset/

Cryptocurrency-Crime Investigation: Fraudulent use of Bitcoin in a Divorce Case

Johnny Botha[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
lleenen@uwc.ac.za

Abstract: Bitcoin and cryptocurrency adoption has increased significantly over the past few years. The significant growth in the industry has been matched by growth of crimes in this domain; not only in scams and dark-web illegal trading, but also in white-collar crimes with fraud and perjury occurring increasingly. With blockchain technology, the world of financial infidelity has become increasingly sophisticated. There is a common belief that blockchain and cryptocurrency provide means of hiding funds from the public or close associates who may not be familiar with the technology. The rise of cryptocurrency has also led to spouses hiding digital assets during divorce settlements. This study presents a use case of a couple in the midst of a divorce where one of the spouses was accused of perjury for failure to declare bitcoin holdings, obtained via Bitcoin mining, and possibly other forms of cryptocurrency and digital assets to the court. The plaintiff is entitled to fifty percent of all assets. While property, stocks, bonds, and bank accounts can easily be traced, cryptocurrency assets are more complex to trace but it is not impossible. This paper illustrates how such a case can be investigated by following the flow of funds on the blockchain, using tools such as Maltego and QLUE. The paper thus presents an investigative process that can be followed for a new category of forensic investigation.

Keywords: Bitcoin, Blockchain, Fraud, Investigation, Perjury

1. Introduction

Bitcoin (BTC), the most popular cryptocurrency on the blockchain with the highest market capital (CoinMarketCap, 2023), has gained wide popularity since it was first created, due to its unique properties, turbulent price swings and surging value. Some may believe bitcoin gives back the power to the people because transactions can be performed without third parties such as banks. However, the popularity of cryptocurrency has attracted various types of fraud and crimes and has become a growing global concern (Trozze, et al., 2022) (Botha, Botha, & Leenen, 2023). Crime in the cryptocurrency space is increasingly moving into the white-collar crime category. Organised criminal groups need some way to "cash out" in ways to evade law enforcement (Gibson Miralis, 2023). The growth of cryptocurrency has also led to cases where spouses hide digital assets during divorce settlements. When a couple, married in community of property, file for divorce, both parties are usually obliged to declare all their assets obtained during their marriage, including digital assets. Most physical assets and bonds are fairly easy to track down, whereas cryptocurrencies could be a bit tougher to trace. A cryptocurrency and forensics investigator may be needed to assist in such cases. Dore (2021) quotes an attorney in New York who noted that awareness among attorneys of these hidden cryptocurrency assets is recent, and that the hardest aspect for attorneys is determine whether such investments exist. This paper illustrates the process that can be followed to identify hidden cryptocurrency assets, and to trace where the funds may have been moved.

Bitcoin is often perceived to offer anonymity. On the contrary, it is traceable. A person's identity is not directly linked to a bitcoin address, but all transactions are public and recorded on the blockchain. Bitcoin and Ethereum are older forms of cryptocurrency and are easier to track down than some of the anonymous or private cryptocurrencies such as Monero (XMR) [(Chainalysis Team, 2023), Zcash and Dash (Hayward, 2021)], but it is not impossible [(Masters Law Group LCC, 2022) (originstamp, 2023)]. These so-called privacy coins make use of varying cryptographic techniques with the aim to obscure details of transactions and to shield users from exposure of related information. Privacy coins have become the preferred method of payment for ransomware demands, criminal transactions on the dark web, and money laundering (Barone & Masciandaro, 2019). These criminals have attracted noteworthy attention from law enforcement and regulators (Hayward, 2021). However, this study's focus is only on tracking the flow of the bitcoins and not any other form of cryptocurrency.

A bitcoin transaction is a digital exchange of cryptocurrencies between two parties. The transaction can occur via an exchange or directly between individuals. All transactions are recorded in a public ledger. This public ledger is called the blockchain. The blockchain is a technology was first revealed by Satoshi Nakamoto in his white paper on 31 October 2008. The technology is the foundation of bitcoin and the heart of all cryptocurrencies (Di Pierro, 2017). When a transaction is recorded, it is also referred to as writing to a block on

the blockchain. Transactions take place without the need for a third party such as a bank or financial institution. The transactions can never be deleted or altered on the blockchain (originstamp, 2023). Blockchain explorers are software applications that allow anyone to view the blocks on the blockchain, as well as addresses and transactions connected to those blocks. Blockchain.com (Blockchain.com, 2023) is one example of a blockchain explorer (originstamp, 2023). To trace a bitcoin transaction, the bitcoin address or transaction-id to be traced has to be entered into the explorer's search field. All the recorded information related to the address or transaction are revealed. Upon entering an address into the blockchain explorer, the explorer will also reveal all transactions made from and to the given address. Other visible information are the block heights at which the transaction was made, as well as the time and date stamps (originstamp, 2023).

Tracing the funds on a blockchain explorer can become a very time-consuming task. Since all the data on the blockchain is publicly available, this falls in the Open-Source Intelligence (OSINT) space. Various OSINT tools exist to assist in crime investigations. One of the most popular tools is Maltego (Maltego, 2023). Maltego has various transforms that can be used to integrate into external data-sources. One of such a data source is the Tatum Blockchain Explorer. Tatum supports over 40 blockchain protocols and over 2000 digital assets (Maltego Technologies, 2023). Maltego makes it much easier to follow the funds on the blockchain. Although Maltego makes it much easier than working on the public blockchain explorers, the tool has limitations. It is almost impossible to match bitcoin addresses with their owners (users). Maltego does not perform clustering of addresses and does not indicate whether an address belongs to a crypto exchange.

One of the tools to perform blockchain investigations is QLUE. This tool was designed for law enforcement and financial investigators to secure evidence of fraud involving crypto. The tool is very effective in following the trace of funds on the blockchain and can reduce time spent on crypto investigations from weeks to merely hours (QLUE, 2023). The tool is capable of clustering certain crypto addresses and can also identify and highlight if an address is flagged as an address known to be fraudulent. The tool can also indicate whether an address is linked to a cryptocurrency exchange or not. For an investigator to identify and prove that transactions linked to an address and then linked to an exchange were executed, is of utmost importance and constitutes valuable information. Exchanges collect personal information on members, which is the type of information of interest to an investigator. However, one cannot request personal information without the involvement of law enforcement. A law enforcement agency can issue a subpoena to an exchange instructing the exchange to reveal the personal information related to a certain crypto or bitcoin address. A subpoena will provide the investigator with personal information, transactional information and evidence that could assist in linking a suspicious cryptocurrency address or transaction to a specific individual. The subpoena will also force the exchange to reveal any other addresses as well as all transactions linked to the target person.

This paper investigates a specific case of possible fraud and perjury during a divorce settlement. Limited information and background have been provided as inputs for the case. With the use of blockchain explorers and OSINT investigating tools such as Maltego and QLUE, it is possible to follow the flow of funds and it could be traced on the blockchain up to a point where a destination address can be reached. From the destination address, further findings and conclusions can be made. It should be noted that the plaintiff has given consent to the researcher to perform this study and to conduct a high-level investigation. Note that no personal information and no blockchain information are exposed in this paper, the bitcoin addresses and transactions have all been sanitised in this paper.

2. Case Background

With the unique properties of bitcoin, it is possible to hide assets from people who are not aware of how the blockchain with its public ledger works or how to trace funds on the blockchain. Tracking down hidden cryptocurrency stashes during a divorce has led to a completely new case category of forensic investigation. If one of the spouses is technically skilled and the other one is not, the prior can easily hide digital assets. The best method to get information on a spouse's cryptocurrency holdings is via a subpoena from a centralized crypto exchange. The alternative is a process that involves a forensic analysis of their computer or other devices to identify a wallet address and then a subsequent blockchain analysis. Crypto asset forensics, cryptocurrency forensics, and blockchain forensics have become an important part of legal divorce cases. Crypto can be stored on exchanges, but it can also be stored on "hot" or "cold" wallets or a combination of both. Hot wallets are connected to the Internet and allow users to easily spend the coins, but they are also more exposed to bad actors. Cold wallets are not connected to the Internet and the private keys are stored on the device itself. Investigating a case where the crypto is stored on cold wallets makes the case a bit more complex than hot

wallets. However, to cash-out one normally, one must go via an exchange or some kind of off-ramp service (Sigalos, 2023).

The use case we consider in this paper, concerns a divorced couple who was married in community of property. The plaintiff is entitled to fifty percent of all their assets. However, the defendant failed to declare any cryptocurrency holdings or assets obtained via bitcoin mining. Bitcoin mining is a method, or the process, of creating new bitcoins by solving complicated mathematical puzzles or problems that verify transactions on the blockchain. When a bitcoin has been mined successfully, the miner receives as a reward a predetermined amount of bitcoin (Baker, 2023). The plaintiff in the divorce case also claimed to have contributed funds to procure fourteen bitcoin Antminers. An Antminer is basically a bitcoin mining computer and is regarded as the most potent in the world when it comes to bitcoin mining. One Antminer could earn 0.1248924 bitcoin monthly (Masa, 2022). Figure 1 is a photo taken of four of the fourteen miners in possession of the defendant. It should be noted that the plaintiff has right of ownership of half of these miners, as well as half of the assets obtained through the mining processes.



Figure 1: Bitcoin Ant-miners

The defendant did not provide any information regarding the bitcoin holdings obtained via the miners. However, the plaintiff managed to obtain two bitcoin addresses linked to the bitcoin mining services. The two addresses given below are the only inputs provided to start this investigation *(please note the full addresses are not disclosed in this paper for privacy reasons):*

- 1DGDY…YSWwM (referred to as BTC_1 in this paper)
- 1JZYv…JPx8A (referred to as BTC_2)

With this limited information, the next step is to follow the funds on the blockchain to determine if the addresses contain any funds, and to view all the transactions performed on these addresses to determine where the funds have been moved to. The aim is to get to a destination address, possibly linked to an exchange. When a destination address has been identified and linked to an exchange, further information could be obtained via subpoena, instructed via an attorney and law enforcement. It should be noted that the destination can be a cryptocurrency wallet address that is not linked to an exchange. For such a case, no personal information will be available to link to the address. An investigator could flag the address and set up a notification when funds move out of the address in the future, with the hopes that it would be moved to an exchange or any other off-ramp where KYC (Know Your Customer) information could be obtained.

The next section illustrates how one could investigate, analyse and follow the funds on the blockchain.

3. Investigation and Findings

Using the blockchain.com explorer and Maltego, it could easily be determined that the two addresses contained 0 bitcoin and the last transaction performed on the addresses were in 2019 (Figure 2 and Figure 3).

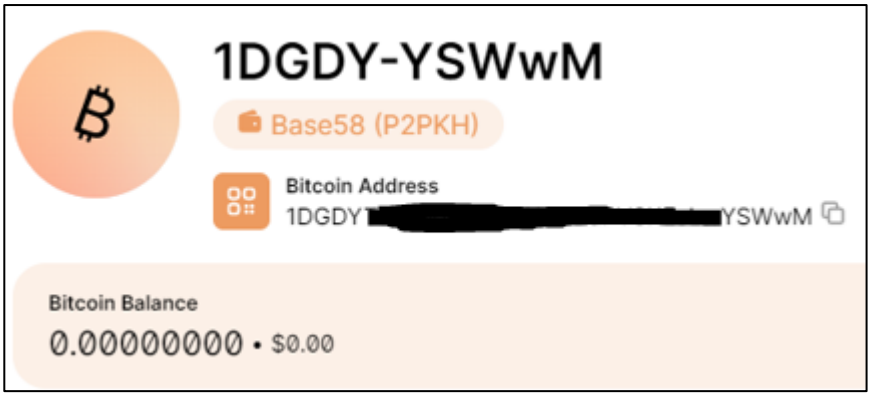


Figure 2: Blockchain.com Explorer - BTC_1 – Balance

The last amount transacted on this address (Figure 4) was 1.07147560 BTC and was sent out of the address into another bitcoin address, namely **177hB…wSxvh**.

Only the first four and last four characters of the address are displayed in Figure 3. If one clicks on the address or copies it from the copy icon next to the "To" address (Figure 3), the full address is revealed.



Figure 3: BTC_1 - Last Transaction Date

At the time of writing, the amount sent from BTC_1 was equivalent to $29, 920.57 (Figure 4).



Figure 4: BTC_1 - Last Transaction Amount

Table 1 lists both input addresses that was received by the plaintiff with their current balance, total received, and total sent.

Table 1: Input Bitcoin Addresses and Balances

BTC Address	Current Balance	Total Received	Total Sent
BTC_1	0	1.07147560	1.07147560
BTC_2	0	0.34892516	0.34892516

Using the Maltego tool, it was discovered that **BTC_1** has 24 incoming transactions and one outgoing transaction to the bitcoin address **177hB…wSxvh** (Figure 5). However, this output address also contained 0 BTC. Further traces have been followed, but all addresses ended up with a 0 BTC balance. **BTC_2** also had several inputs from various address and transactions and one output to the address **17rDj…wtCbK** (Figure 6). Similar to **BTC_1**, the output address had a balance of 0 BTC. The trace has also been followed and all addresses had a balance of 0 BTC. Since the transaction dates are very old, 2019, the assumption can be made that 0 bitcoins will be left in any of the addresses and that the defendant has already cashed out via an offramp such as an exchange. An offramp is a company or service that allows a client to convert cryptocurrency back in fiat money (Rockwallet, 2023). The audit trail could still be followed to a destination address to indicate how many bitcoins were transacted. If the defendant has cashed out the bitcoins, the plaintiff is entitled to a payment in cash equivalent to the amount of these bitcoins.

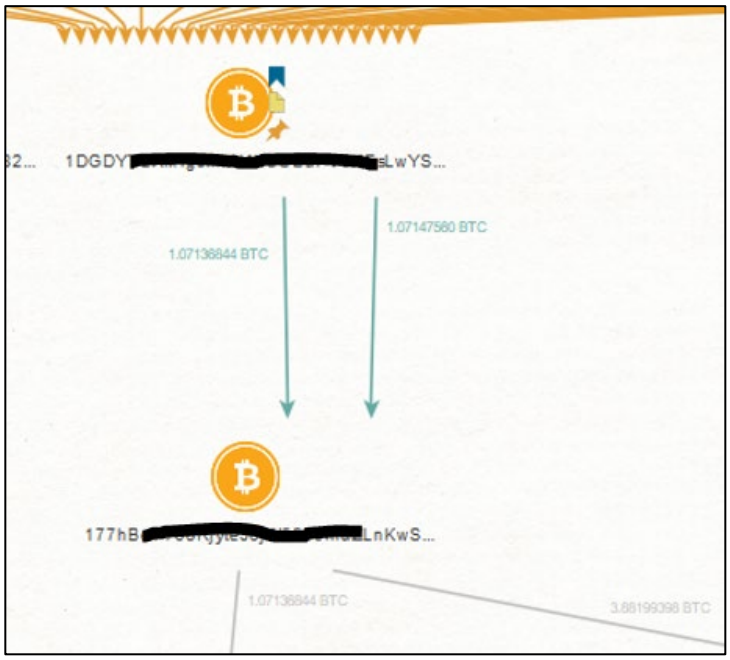


Figure 5: BTC_1 – Output Transaction

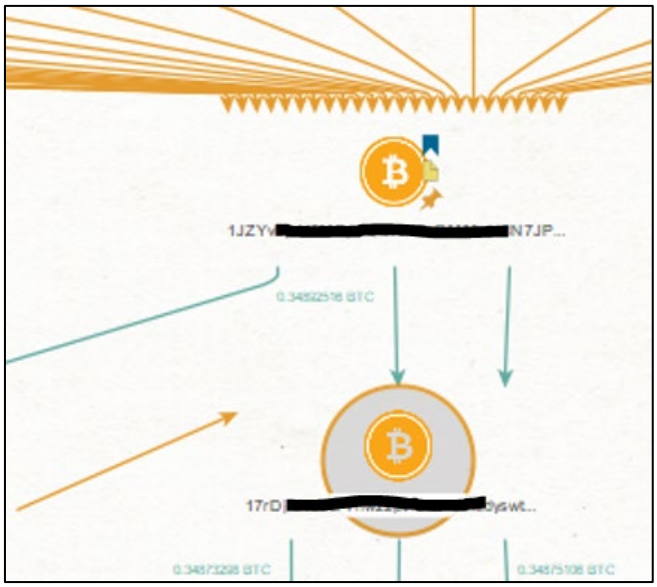


Figure 6: BTC_2 - Output Transaction

After several attempts to follow the funds, no clear conclusion or findings could be reached. The researcher reached out to a crypto-crime investigator who has assisted in a previous research study (Botha, Pederson, & Leenen, 2023). The investigator is involved in big cases such as the Mirror Trading International (MTI) scam as well as the Finalmente Global scam. MTI was a network or multi-level marketing scheme that claimed to offer automated cryptocurrency trading services via bots (Vermeulen, 2022a). Finalmente Global has claimed to take bitcoin investments from members to buy online advertising space. Both platforms have been shut down and are under investigation (Vermeulen, 2022b).

Access to the cryptocurrency graphing and investigation tool QLUE is restricted to licensed versions. Through collaboration with the crypto-crime investigator, access to the tool was received and the investigation was continued. QLUE makes it easier to follow the funds on the blockchain and it gives the investigator clustering capabilities. It can cluster addresses belonging to the same user, or addresses linked to a certain exchange. It also allows for the investigator to flag certain addresses as fraudulent or indicates cases where an address has been linked to a known scam. For the purpose of this investigation, a link to an exchange is required. This will allow the investigator, through an attorney, to send a subpoena to the exchange. A subpoena instructing the

exchange to reveal the personal information as well as indicating if any cash-out to a bank was performed, will then be issued. If a cash-out has been done, the cash-out transaction details need to be provided to the investigator. The type of information required is the amount that was cashed out as well as the dates and bank details.

The same two addresses, **BTC_1** and **BTC_2**, have been provided to TCG Forensics as inputs. By using QLUE, it was discovered that the transactions from both addresses ended up in an address that belongs to Luno (Luno, 2023), a South African cryptocurrency exchange (Figure 9). Figure 7, Figure 8 and Figure 9 show the flow of funds from the input addresses up to the destination address that is linked to Luno. In addition to the two provided input addresses, another input address, **1NTQf…8Es7e**, was also revealed (Figure 7). The address appears to be from the same source mining activity based on the clustering done by QLUE. This address will be referred to as **BTC_3**.

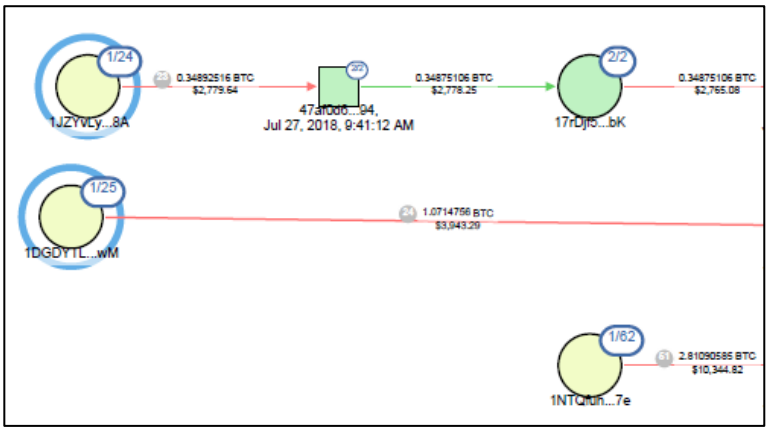


Figure 7: Inputs into QLUE (Flow of funds - a)

By continuing to follow the funds, three addresses have been identified with amounts sent to the destination address into Luno (Figure 8). QLUE hides the addresses in the diagram. It should be noted that bitcoin addresses are visualised in QLUE as round circles and transactions as square blocks. An arrow going out of a square block (transaction) into a circle (address) in Figures 7, 8 and 9 refers to an incoming flow of funds into a bitcoin address; an arrow going out of a circle (address) into a square block (transaction) indicates an outgoing flow of funds. An incoming flow of funds will always go from a transaction into an address and an outgoing flow of funds from an address into a transaction.

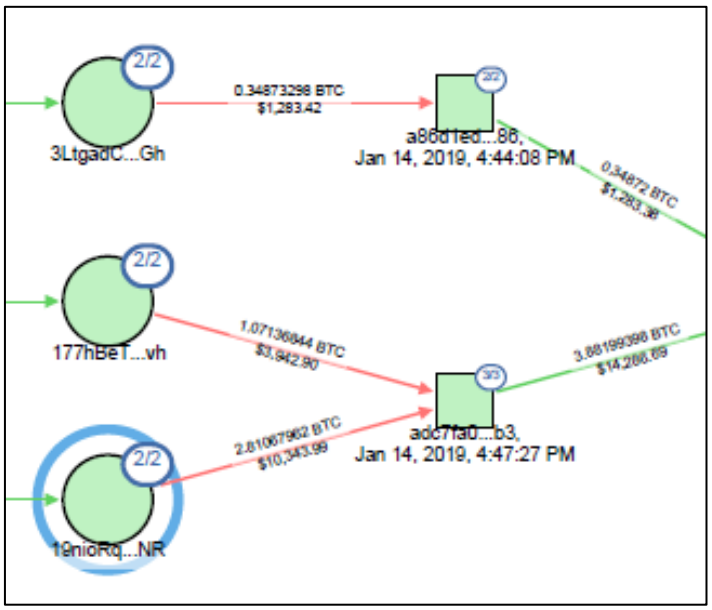


Figure 8: Flow of funds (continue - b)

Table 2 lists the three addresses (Figure 9) with the amounts sent from each address into the destination address, **1DRfj…xY9KN**, (Figure 10). The first column indicates the origin or starting point of the flow of funds,

the second column the last address before the funds moved to the destination address and the third column lists the amount sent to the destination address via each input stream (Table 2).

Table 2: Last Addresses Before Destination Address

From Input Address	Last Address Before Destination	Total Sent
BTC_1	3Ltga…a2eGh	0.34873298
BTC_2	177hB…wSxvh	1.07136644
BTC_3	19nio…WNUNR	2.81007962

The sum of these three amounts in Table 2 is equal to 3.916319358 BTC. Looking at the outflow of funds in Figure 9, only 3.88199398 BTC has been sent out of Luno to another address external from Luno. Thus, the remainder of 0.34872 BTC that went into the destination address could possibly still be in the balance.

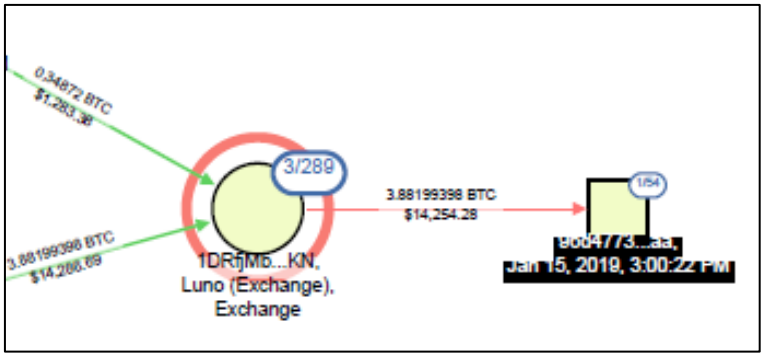


Figure 9: Flow of funds (continue - c) – Destination Address

However, QLUE indicated that the current balance on the destination address (Figure 10) shows a balance of 0 BTC. This means the full amount of BTC was transferred out of the destination address.



Figure 10: Destination Address

Upon further investigations using QLUE, it has been discovered that the remaining amount has been transferred from the Luno exchange into another address outside of Luno (Figure 11). The address it was sent to was also flagged by QLUE as a Hot wallet, and linked to the cryptocurrency exchange Bitssa (Bitssa, 2023), (see Figure 12). To see if this transaction was performed by the defendant one needs to obtain a transaction list from Luno. This can be obtained with a subpoena issued by law enforcement. At the time of writing this paper, no subpoena had been requested to take the investigation further. The aim of this paper is to reveal the amounts of bitcoin the defendant has been hiding from the spouse in the divorce settlement by following certain steps; subsequent legal action is beyond the scope of our research.

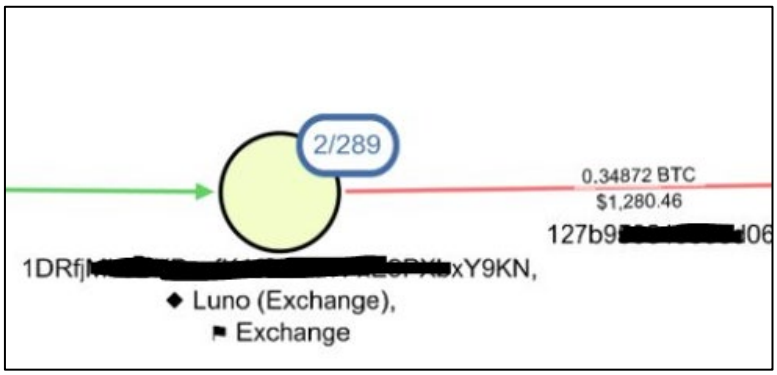


Figure 11: Remainder Amount - Transferred from Luno to Bitssa.in

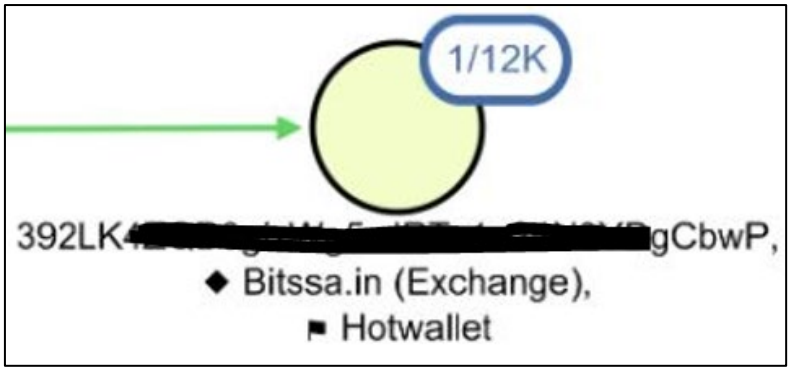


Figure 12: Bitssa.in – Hot wallet Address

It should be noted that the findings in this study were only linked to the two input addresses provided by the plaintiff. It is also possible that the defendant hid more cryptocurrency investments obtained by other means during the marriage. These assets should also be declared by the defendant to the court during the divorce settlement case. However, this is outside the scope of this investigation.

This study and investigation illustrated that it is not impossible to trace funds on the blockchain. Various tools exist to assist in following funds and movements on the blockchain. With an investigation on the blockchain, the aim is to find links to an exchange. An exchange normally stores KYC information of members. The idea is to obtain personal information that could be linked to certain bitcoin addresses and transactions. Once personal information can be identified, the investigator could attempt to contact the person of interest. If no response is received, alternative OSINT investigation techniques will have to be followed to locate the person.

4. Conclusion

This paper gives an overview and introduction to white collar crime in the crypto space. Many people believe that crypto is anonymous and that they can use the technology to hide some of their assets from people such as a spouse, for example. However, as illustrated in this paper, that can be contrary to belief as every transaction on the blockchain is transparent and visible to the public. When one uses blockchain explorers and investigation tools such as Maltego or QLUE, it is possible to trace the flow of funds on the blockchain. An investigator simply needs a starting point to follow the funds to a destination that is linked to an exchange. Most exchanges collect personal information on all members. An investigator can issue a subpoena via an attorney to an exchange. This will instruct the exchange to reveal the required personal information on a target person linked to a crypto address within their platform.

This paper describes an investigative approach to a divorce settlement case. The defendant was accused of not declaring digital assets in the divorce settlement. The defendant had obtained bitcoins via bitcoin mining processes during their marriage. These mining computers had also been procured during the time they were still married. Two bitcoin addresses have been provided by the plaintiff as inputs and a starting point on the blockchain. From the high-level investigation on the flow of funds, it became clear that the addresses provided as inputs by the plaintiff showed a balance of 0 BTC. Further findings were that 3.916319358 BTC had been sent to a destination address that is linked to the Luno exchange. Only 3.88199398 BTC had left the Luno exchange, and these funds appear to have been cashed out. The remainder had been sent to another exchange. To get the proof of the cash-out and the remainder been sent to another exchange by the defendant, a subpoena is required to instruct the exchange to reveal more information such as all transaction details of the defendant. Thus, based on these findings, there is evidence that the accused obtained 3.916319358 BTC from the mining services while being married. The bitcoins had been hidden from the plaintiff during the divorce settlement. Half of value of these bitcoins should be paid out to the plaintiff.

Acknowledgement

The authors thank Thor Pederson from TCG Forensics for his continued support and advice.

References

Baker, B. (2023, Mar 27). *What is Bitcoin mining and how does it work?* Retrieved from www.bankrate.com: https://www.bankrate.com/investing/what-is-bitcoin-mining/

Barone, R., & Masciandaro, D. (2019). Cryptocurrency or usury? Crime and alternative money laundering techniques. *Eur J Law Econ 47* (pp. 233–254). Springer. doi:https://doi.org/10.1007/s10657-019-09609-6

Bitssa. (2023, Aug 11). *Bitssa Swap*. Retrieved from https://swap.bitssa.com/: https://swap.bitssa.com/

Blockchain.com. (2023, Aug 11). *Blockchain.com Explorer*. Retrieved from www.blockchain.com: https://www.blockchain.com/explorer

Botha, J., Botha, D., & Leenen, L. (2023). An Analysis of Crypto Scams during the Covid-19 Pandemic: 2020-2022. *18th International Conference on Cyber Warfare and Security* (pp. 36-48). Towson University, Baltimore County Maryland, USA: Academic Conferences International Limited.

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

Chainalysis Team. (2023, May 4). *Monero: All About the Top Privacy Coin*. Retrieved from www.chainalysis.com: https://blog.chainalysis.com/reports/all-about-monero/#:~:text=Monero%27s%20privacy%2Denhancing%20features,-Monero%27s%20primary%20purpose&text=The%20Monero%20blockchain%20employs%20diverse,user%20generated%20a%20given%20signature.

CoinMarketCap. (2023, Aug 11). *Home Page*. Retrieved from https://coinmarketcap.com/: https://coinmarketcap.com/

Di Pierro, M. (2017). What is Blockchain. *Computing Prescriptions*. Retrieved from https://cse.sc.edu/~mgv/csce190f18/diPierro_mcs2017050092.pdf

Dore, K. (2021, Jun 1). *Spouses in divorce proceedings are using cryptocurrency to hide money. Here's how experts find it*. Retrieved from www.cnbc.com: https://www.cnbc.com/2021/06/01/divorcing-spouses-are-using-cryptocurrency-to-hide-money-how-experts-find-it.html

Gibson Miralis, N. (2023, Aug 11). *How is crypto crime becoming increasingly white collar?* Retrieved from www.lexology.com: https://www.lexology.com/library/detail.aspx?g=453fc875-be12-4701-b3ed-585aad9a2191

Hayward, A. (2021, May 31). *What Are Privacy Coins? Monero, Zcash, and Dash Explained*. Retrieved from https://decrypt.co: https://decrypt.co/resources/what-are-privacy-coins-monero-zcash-and-dash-explained

Luno. (2023, Aug 11). *Luno Home Page*. Retrieved from https://www.luno.com/en/za: https://www.luno.com/en/za

Maltego. (2023, Aug 11). *Maltego Home Page*. Retrieved from Maltego.com: https://www.maltego.com/

Maltego Technologies. (2023, Aug 11). *Tatum Blockchain Explorer*. Retrieved from www.maltego.com: https://www.maltego.com/transform-hub/tatum-blockchain-explorer/

Masa, W. (2022, Nov 24). *What Is Antminer? How Much Does A Antminer Make A Day?* Retrieved from https://bitkan.com: https://bitkan.com/learn/what-is-antminer-how-much-does-a-antminer-make-a-day-8792

Masters Law Group LCC. (2022, Nov 4). *CAN HIDING CRYPTO FROM YOUR EX GET YOU IN LEGAL TROUBLE?* Retrieved from www.masters-lawgroup.com: https://www.masters-lawgroup.com/cryptocurrency-and-divorce/can-hiding-crypto-from-your-ex-get-you-in-legal-trouble/

originstamp. (2023, Nov 11). *How to Trace Bitcoin Transactions [Full Guide]*. Retrieved from www.originstamp.com: https://originstamp.com/blog/how-to-trace-bitcoin-transactions/

QLUE. (2023, Aug 11). *QLUE Home Page*. Retrieved from www.qlue.io: https://qlue.io

Rockwallet. (2023, Aug 11). *Off-Ramp and On-Ramp Crypto, What You Need To Know*. Retrieved from Rockwallet.com: https://www.rockwallet.com/blog/off-ramp-and-on-ramp-crypto-what-you-need-to-know#:~:text=On%20the%20other%20hand%2C%20off,ramps%20allow%20you%20to%20exit.t=On%20the%20other%20hand%2C%20off,ramps%20allow%20you%20to%20exit.

Sigalos, M. (2023, May 20). *A husband hid $500,000 in bitcoin during a divorce — and got busted by a crypto hunter*. Retrieved from https://www.cnbc.com: https://www.cnbc.com/2023/05/20/bitcoin-in-divorce-how-spouses-hide-assets-crypto-hunters-find-them.html

TCG Forensics. (2023, Aug 11). *Most Prominent Investigations Into Fraud, Cybercrime And Violent Crimes*. Retrieved from www.tcgforensics.co.za: https://www.tcgforensics.co.za/

Trozze, A., Kamps, J., Akartuna, E., Hetzel, F., Kleinberg, B., Davies, T., & Johnson, S. (2022). Cryptocurrencies and future financial crime. *Crime Sci 11.* Springer. doi:https://doi.org/10.1186/s40163-021-00163-8

Vermeulen, J. (2022a, Apr 12). *Good news for Mirror Trading International scam victims*. Retrieved from https://mybroadband.co.za: https://mybroadband.co.za/news/cryptocurrency/440928-good-news-for-mirror-trading-international-scam-victims.html

Vermeulen, J. (2022b, Jan 7). *Finalmente Global shuts down, blames death of Mirror Trading International*. Retrieved from https://mybroadband.co.za: https://mybroadband.co.za/news/cryptocurrency/381646-finalmente-global-shuts-down-blames-death-of-mirror-trading-international.html

Capture the Flag with ChatGPT: Security Testing with AI ChatBots

Ellis Casey and David Chamberlain
Airbus, Newport, UK

ellis.casey@airbus.com
david.chamberlain@airbus.com

Abstract: Penetration testing, commonly referred to as pen testing, is a process of assessing the security of a computer system or network by simulating an attack from an external or internal threat actor. One type of pen testing exercise that has become popular among cybersecurity enthusiasts is called Capture the Flag (CTF). This involves solving a series of challenges that simulate real-world hacking scenarios, with the goal of capturing a flag that represents a piece of sensitive information. Recently, there has been a growing interest in the use of natural language processing (NLP) and machine learning (ML) technologies for penetration testing and CTF exercises. One such technology that has received significant attention is ChatGPT, a large language model (LLM) trained by OpenAI based on the GPT-3.5 architecture. The use of ChatGPT in CTFs has several potential benefits for participants and organisers, including more dynamic and realistic scenarios and enhanced learning experiences, and enhance the effectiveness and realism of CTFs.. Future research can explore more sophisticated models and evaluate the effectiveness of ChatGPT in improving the performance of participants in CTFs.

Keywords: ChatGPT, Chatbot, CTF, Penetration testing, Hacking

1. Introduction

In recent years, cyber-attacks are becoming more frequent and sophisticated. To keep pace with these evolving threats, it is important to continually improve and innovate the tools and techniques used in cybersecurity. One promising area of innovation is the use of artificial intelligence (AI) in cybersecurity, particularly in the area of penetration testing. In this essay, the use of ChatGPT, a large language model trained by OpenAI, will be explored in conducting penetration testing on Capture the Flag (CTF) exercises.

CTF is a popular cybersecurity competition that simulates real-world scenarios for participants to test their skills in offensive and defensive security. CTFs involve a variety of challenges, ranging from web application exploits to network and system vulnerabilities. One of the main objectives of CTFs is to provide a safe and controlled environment for participants to learn about cybersecurity concepts and techniques, and to practice penetration testing skills.

Penetration testing and CTF exercises are two critical areas of cybersecurity that require constant innovation to stay ahead of evolving threats. One such innovation that has recently garnered attention in the field is the use of natural language processing (NLP) and machine learning (ML) technologies to enhance the effectiveness of these exercises. Specifically, ChatGPT, a large language model trained by OpenAI, has shown promise in automating the process of generating attack vectors and testing them against target systems in CTFs. This essay will also explore how to leverage ChatGPT when conducting CTF exercises, including technical requirements and integration functions. Examples will be discussed on how ChatGPT and other chatbots can be leveraged in CTFs.

ChatGPT can be used to automate the process of generating attack vectors and testing them against a target system or network. By leveraging its ability to understand and generate natural language, ChatGPT can craft attack scenarios that are more sophisticated and realistic than traditional approaches. This makes it a powerful tool for conducting CTF exercises, as it can generate challenges that are tailored to the specific strengths and weaknesses of the participants.

Using ChatGPT for CTF exercises has the potential to significantly enhance the realism and effectiveness of these exercises. In addition to generating challenges, ChatGPT can also act as a virtual adversary, providing hints and feedback to participants based on their progress. This can help to create a more engaging and dynamic CTF experience, while also providing valuable feedback on participants' strengths and weaknesses, helping them overcome obstacles and learn new skills.

However, the use of ChatGPT in pen testing and CTF exercises also raises important ethical and legal questions. The technology can potentially be used to automate and scale attacks against real-world targets, which could have serious consequences for individuals and organisations. As such, it is important to consider the ethical implications of using ChatGPT for these purposes and to ensure that its use is in line with established ethical guidelines and legal frameworks.

2. Leveraging ChatGPT in CTFs

ChatGPT and other chatbots can be leveraged in CTFs to enhance the realism and effectiveness of the exercises. For example, ChatGPT can be used to create dynamic challenges that respond to participants' actions and progress. This can help to create a more engaging and realistic CTF experience, while also providing valuable feedback on participants' strengths and weaknesses, providing hints and suggestions for participants who are struggling with a challenge, helping them overcome obstacles and learn new skills.

In addition, ChatGPT can be used to simulate the behaviour of real-world threat actors by generating attack scenarios that mimic their tactics, techniques, and procedures (TTPs). This can help participants to develop a better understanding of how threat actors operate and how to defend against them.

Further ways to leverage ChatGPT in CTFs is as a tool to teach participants about cybersecurity concepts and best practices. For example, ChatGPT can be used to generate interactive tutorials on topics such as password cracking, buffer overflows, and SQL injection. Providing participants with hands-on experience, they can develop a deeper understanding of cybersecurity concepts and how to apply them in practice.

The use of ChatGPT in CTFs has several potential benefits for participants and organisers. One of the main advantages is that ChatGPT can provide more dynamic and realistic scenarios that simulate real-world threats. ChatGPT can generate responses based on the participants' solutions and adapt to their level of knowledge and experience, making the challenges more personalised and engaging.

3. Research Plan

Within this research, each step has been manually entered to ensure validity of the results in order to demonstrate ChatGPT's practical usage in security testing within cybersecurity, as well as measuring its success rate in completing various tasks from user prompts.

What is being studied

- Using ChatGPT for assistance in CTF challenges from Hack the Box (HTB), which is a gamified cybersecurity learning platform
- Assessing its collected training set of cybersecurity concepts, tools and techniques
- Evaluating its ability to process context of the challenges and provide accurate solutions or hints where possible
- Identifying limitations of ChatGPT's capabilities, such as potential for misuse, and areas where human intervention is still necessary
- Is a level of fundamental knowledge of cybersecurity required, in order to interrogate and elicit an appropriate response from ChatGPT

Potential output of this research

- Gain an understanding of ChatGPT's capabilities and limitations in the context of cybersecurity, and how it could be integrated into security testing workflows
- Establishing a performance benchmark in ChatGPT's current state
- Identifying areas of improvement for the application
- Gain a better understanding of how crafting the commands used will determine the validity or usefulness of the output.

Considerations

- ChatGPT struggles with identifying and assisting with newer CVEs and exploits
- When queried about a newer CVE, it will pretend to know what it is and 'hallucinate' incorrect information about its usage. These hallucinations refer to the chatbot confidently generating seemingly realistic responses that do not coincide with real-world scientific knowledge. Since ChatGPT is an example of an LLM (Large Language Model), it has a tendency to falsify information and present it as factual. When a user sends a prompt to ChatGPT, it gets broken down into small parts called tokens, which could be as short as a single character or word. It then processes these tokens, using patterns it learned during its training to predict what word (token) will come next in the sequence. Each token iteration is assigned a probability and the token with the highest probability is chosen to continue the text generation. This process is repeated until the full response has been generated.

- This makes the response seem human-like, in the sense that it doesn't like to say "I don't know" when it doesn't know
- Recent update - Instead of substituting the CVE the user provides by adding it to the end of a CVE database URL and hallucinating information, it now seems to interpret the date of the CVE given and tell the user it only has information up to September 2021. This change was observed between approximately the 9th February 2023 update to the 23rd March 2023 update.
- It states the last known CVE it knows is Log4Shell
- This makes it easier for security testers to know what information is accurate, however adding a confidence scale will ultimately be the best solution for ensuring accuracy, along with providing a source(s) for the output it provides

4. Selected HTB Machines

Machines were chosen with generally high user ratings in order to try and give ChatGPT what is perceived as a greater chance of success. It was also decided to exclusively use Linux machines for consistency and used a mix of easy and medium challenges involving a variety of vulnerabilities, in order to vary the potential difficulty for the chat bot. The overall thoughts on ChatGPT's usefulness and limitations in certain areas of each machine have been recorded. 'Boot-to-root machines' were selected in order to simulate a real life scenario, as opposed to short challenges. Boot-to-root challenges involve going from booting the machine to escalating to 'root' or 'admin' level privileges on a vulnerable system. It's worth noting that during the testing phase, most of the machines chosen were active, meaning that there were no online write-ups for these challenges at the time.

The following section contains the description and general steps taken during each challenge by HTB.

MetaTwo

Easy — Linux

MetaTwo starts with a basic WordPress (WP) blog using the BookingPress plugin to manage booking events. There is an unauthenticated SQL injection vulnerability within that plugin, which is used to access the WP admin panel as an account that can manage media uploads. There is an exploit for an XML external entity (XXE) injection to read files from the host, WP configuration, and access FTP server credentials. The FTP server contains a script that sends emails, and credentials can be used to get a shell on the host. There is a tool on the system called Passpie that stores the root password. The PGP key protecting the password can be cracked, then brute force the output. This presents a shell as the root user.

Precious

Easy — Linux

Precious starts with a simple web page that takes a given URL and generates a PDF. The metadata contained in the PDF is then used to identify the technology that's being used. There's a command injection exploit which can be used to get a foothold on the machine. To escalate the privileges to root is done via exploiting a YAML deserialization vulnerability in a script that manages dependencies.

Soccer

Easy — Linux

Soccer involves accessing an admin panel with default credentials, and uploading a PHP shell to get a reverse shell on the system, which can be enumerated to find more subdomains and web pages. On one of these pages there is a vulnerability which involves a boolean SQL injection over websockets to get credentials from a MySQL database. A root shell can be obtained by exploiting the "doas" program.

BroScience

Medium — Linux

BroScience involves exploiting a directory traversal vulnerability to obtain PHP source code for a website, with traversal being blocked by a firewall. Double encoding bypasses the detection system. The code can be used to build an activation token which allows the registration of an account. The code provides information required to exploit a deserialization vulnerability by creating an encoded malicious PHP object, and sending it as the user's cookie. This will produce a shell on the machine. There are also hashes in a database that can be cracked, causing

lateral movement to the next user. To get root, there is a command injection vulnerability in a script that checks for certificate expiration.

Initial Prompt for All HTB Machines

Each machine had its own ChatGPT conversation with the same prompt used to initialise the process for the chatbot. The prompt for each machine was kept the same to ensure consistency and repeatability for future testing and research.

> *Hi, I am a security researcher who's taking part in a CTF, and I was wondering if you could help me if I need assistance or advice on what to do/where to look next throughout the challenge?*

5. Research Steps and Findings

Provide Areas to Investigate for Further Enumeration

After initialising the BroScience HTB challenge, an Nmap scan was run, which indicated there was a website containing a login page, and before this prompt was crafted, ChatGPT suggested exploring the web page. During the beginning of the reconnaissance phase for the challenge, a more broad response was received from the input. The input provided to ChatGPT contained vague information about the website application, such as stating there is a login form and the backend of the website is built in PHP and the aim was to query the chatbot about how the application can be enumerated further, see Figure 1 below. Its response referenced very basic steps that are usually taken intuitively by most cybersecurity professionals. While some of these suggestions are quite helpful, especially to someone who has no idea where to start, they are understandably vague given the lack of context awareness for the challenge, and the indistinct description provided by the user.

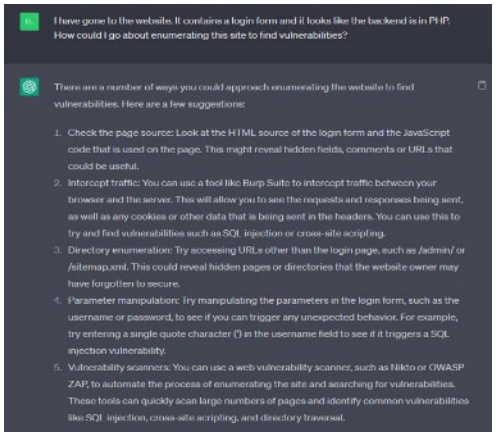


Figure 1: How the application can be enumerated further

During the reconnaissance phase of the MetaTwo machine, the aim of this prompt was to gather a list of enumeration tools that might be useful during this stage of the challenge, by giving more context of what was on the machine as well as specific services that are running. Prior to this, the only information provided to ChatGPT was the output of an Nmap scan. This time a better list of enumeration tools was received by being specific and emphasising potentially important information about the machine. The main difference between this prompt shown in Figure 2, and the previous one shown in Figure 1, was that the user was more specific in identifying what services and tools were being used. It suggested a vulnerability scanner called 'WPScan' based on the type of web management system utilising WordPress, as well as certain tools like 'Hydra' to potentially brute-force login credentials for the FTP server. These suggestions are mixed in with other common tools like Metasploit and Nmap, which was interesting since it had already been provided with a scan of any network services using Nmap previously.

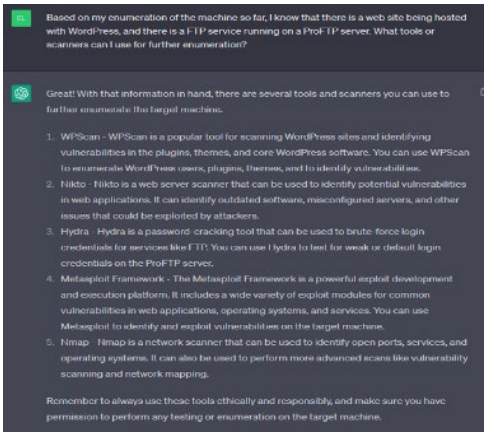


Figure 2: The main difference between this prompt

Since enumeration is the most essential part of any penetration test or red team engagement, it's imperative that ChatGPT provides useful information during this phase. Due to the chatbot not having access to any more information about the machine than the user provides, it lacks any sort of context awareness regarding the challenge, and therefore lacks any nuances and specifics of the scenario other than what information it's fed. This also means it has the potential to produce incorrect suggestions, and while it can provide general suggestions based on common practices, it will not be able to distinguish more effective and less effective techniques in specific situations.

In its current state, it's not advisable to heavily rely on ChatGPT to suggest areas for enumeration due to its lack of context awareness and deep technical knowledge within specific domains of cybersecurity. Its behaviour and output are shaped by its training set. It can, however, definitely prove useful for someone who is starting out in cybersecurity or stuck during a penetration test.

Suggesting Attack Vector and Bypass Detection System

ChatGPT provided a seemingly correct exploit for this stage of the BroScience machine, and gave an example to demonstrate along with a high level explanation of how the exploit works, and how to know if the target machine is vulnerable to this type of attack. However, it did take a bit of back and forth with the chatbot to get to this point, having previously gone down rabbit holes that led to dead-ends by suggesting that a login form was vulnerable to SQL injection or brute forcing credentials. It also previously suggested using several web vulnerability scanners such as Nikto and Dirbuster, which did find some hidden directories, and led to discovering the missing path parameter.

ChatGPT provided some good suggestions for this vulnerability, as it noted there was a detection system in place, and double URL encoding worked in this particular case. However, it did not give sufficient detail on how to encode the payload or give any more specific examples, only a potential direction to explore.

ChatGPT performed moderately well overall in suggesting attack vectors. However, suggesting this particular exploit was after a lot of back and forth with the chatbot, to get to the point where sufficient information was provided to ChatGPT during the enumeration phase of the challenge. Once it was confirmed the attack vector works, the chatbot was able to provide suggestions for bypassing the detection system. Demonstrating that proving enough information during enumeration is paramount for ChatGPT to suggest accurate vulnerabilities, areas for exploitation, and reduces the potential for incorrect suggestions.

Analysing and Deobfuscating Code, and Decoding Base64

With some PHP files retrieved throughout the BroScience machine by injecting the website with an encoded local file inclusion (LFI) path, ChatGPT was particularly useful at analysing the code and describing how it works. After giving the contents of one of the PHP files, ChatGPT responds with a description of what the code does and a high level overview of how the script works. ChatGPT shows promising utility for explaining complex code, by explaining its logic and what each function does.

One area where ChatGPT proves incredibly useful is code de-obfuscation. In this case, PowerShell de-obfuscation. Of course, there are some online tools out there that can do this already, but given the extra explanation ChatGPT provides, it is the preferred tool on the market.

For the MetaTwo machine, ChatGPT had previously suggested I use WPScan to enumerate the website since it was using WordPress, and it provided suggestions based on the output of this scan. It was discovered that an old, vulnerable version of the BookingPress plugin was being used. After asking ChatGPT if there are any known exploits for that version, it states that it cannot provide any known exploits as it goes against OpenAI's usage policy, which will be discussed further later in this paper. After finding a working exploit online, it allowed the retrieval of the /etc/passwd directory which was Base64 encoded. After asking ChatGPT to decode this, it misses a crucial part of the CTF in this output, as there is no 'jnelson' user displayed in ChatGPT's result. Decoding it via the command-line to compare results shows that there are some differences in the way both tools decode it.

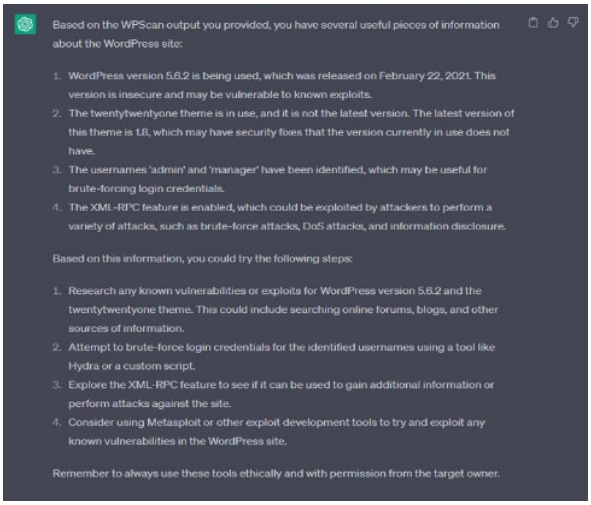
After a short back-and-forth with ChatGPT, it states there is no user 'jnelson' in the encoded text it was provided shown in Figure 12, and was then asked to decode it again. After decoding it again, the chatbot outputs the expected user account with a minor difference - denoting the full name of the user as 'John Nelson', which is not present in the output of the command-line Base64 decoding. To confirm this was a hallucination, it was told there was another missing user and asked it to decode it again. This time, ChatGPT added 'jsmith' as a new user, which confirms that ChatGPT might not be reliable for decoding various encoding formats, which can be seen in Figure 13.

Something interesting to note is that ChatGPT redacts the password for a database, as well as removing the database charset rules and collate type by filtering out potentially sensitive information. It also misses out critical FTP information, such as the username and password. The password in particular is crucial to the next stages of the MetaTwo CTF challenge. Ultimately it is useful to decode Base64 quickly, although it's best not to rely on it getting the syntax 100% accurate, and would instead prefer to use CyberChef or stick to decoding it via the terminal. When asking the chatbot to display the password, it refuses for ethical concerns, no matter how the prompt was phrased.

In regards to code interpretation, ChatGPT does very well. However, it has a tendency to pretend it knows what it's saying and can hallucinate information. During the decoding process, it seems to filter and redact certain information it perceives as sensitive or exploitable. Bypassing the filter to reveal such sensitive information will be covered later in the paper.

Interpreting Scan Results

After providing the output of the WPScan for the MetaTwo machine to ChatGPT, it noticed that an outdated version of WordPress was being used, along with an outdated theme version in Figure 3. It also identified two user accounts "admin" and "manager" and suggested the idea of brute-forcing these credentials. On top of picking out key information from the scan, the chatbot suggested some steps to take based on the information it found. The most useful suggestion it gave in this case was to research vulnerabilities for this outdated software. This was conducted after ChatGPT was provided an Nmap output and told about various services running on the machine, such as FTP on a ProFTP server. The chatbot had also provided the exact command for WPScan to enumerate users, plugins, themes, and so on.


Based on the WPScan output you provided, you have several useful pieces of information about the WordPress site:

1. WordPress version 5.6.2 is being used, which was released on February 22, 2021. This version is insecure and may be vulnerable to known exploits.
2. The twentytwentyone theme is in use, and it is not the latest version. The latest version of this theme is 1.8, which may have security fixes that the version currently in use does not have.
3. The usernames 'admin' and 'manager' have been identified, which may be useful for brute-forcing login credentials.
4. The XML-RPC feature is enabled, which could be exploited by attackers to perform a variety of attacks, such as brute-force attacks, DoS attacks, and information disclosure.

Based on this information, you could try the following steps:

1. Research any known vulnerabilities or exploits for WordPress version 5.6.2 and the twentytwentyone theme. This could include searching online forums, blogs, and other sources of information.
2. Attempt to brute-force login credentials for the identified usernames using a tool like Hydra or a custom script.
3. Explore the XML-RPC feature to see if it can be used to gain additional information or perform attacks against the site.
4. Consider using Metasploit or other exploit development tools to try and exploit any known vulnerabilities in the WordPress site.

Remember to always use these tools ethically and with permission from the target owner.

Figure 3: An outdated theme version

ChatGPT was also queried about Nikto scan results in Figure 4 for MetaTwo to see if it would offer up any more suggestions on top of the previous ones. It provided a lot of potential areas to explore based on the output of

the scan, which produced more tailored guidance on what to look into next. This step was conducted immediately after analysing the WPScan output.

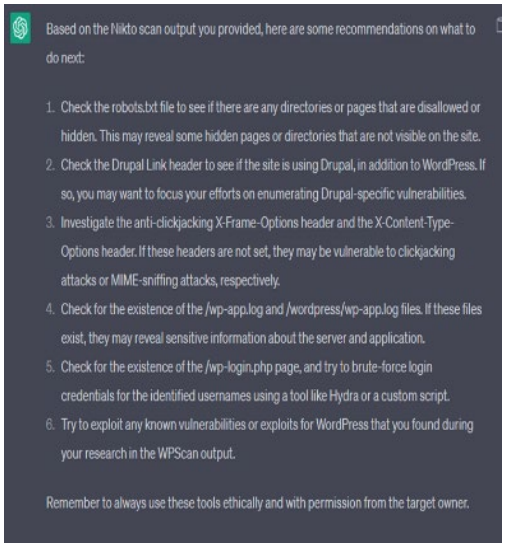


Figure 4: Nikto scan results

After providing ChatGPT with MetaTwo's Nmap scan results in Figure 5, it gives an overview of results, noting any important information that might be required to progress in the CTF, such as open ports, running services and domains. Interestingly it then suggested that the next step should be to run a network scanner like Nmap to identify what services are running on the network, which seemed redundant.

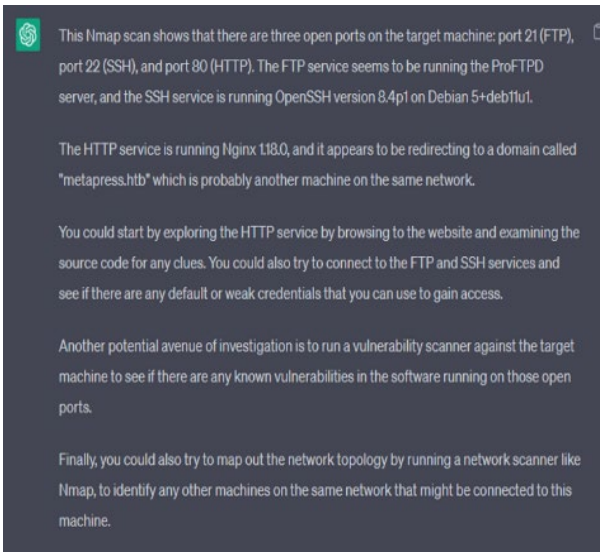


Figure 5: MetaTwo's Nmap scan results

ChatGPT is particularly useful for attaining a high level overview of scan results, and is especially useful if the output of the scan is very long and tedious to go through line by line. By providing ChatGPT with scan results, it can help understand and pick out what parts are worth investigating further which can make the penetration test process much more efficient, particularly for those who are inexperienced and might not know what to look out for. These scan results can lead to the chatbot assisting in identifying attack vectors by analysing potential vulnerabilities in scan results.

While ChatGPT has general knowledge about security standard security tools mentioned in its training set, it doesn't necessarily possess deep expertise in specific features or intricacies of every tool. Therefore its guidance on advanced usage in certain tools may be limited. The analysis will be based solely on the information provided in the scan results, as well as any other information the user has shared with the chatbot. This means that it may not have an understanding of the wider context of the target system, and therefore it's crucial to consider

utilising human expertise in looking over the output of the scan, in case there is anything that ChatGPT could've missed.

Hallucinations

Some time after the 9th February 2023 update, in Figure 6 ChatGPT was prompted to provide information about a CVE used in HTB's MetaTwo machine which was reported in 2022, to see how it would respond. It gave completely incorrect information about the CVE, stating it's a remote code execution (RCE) vulnerability for certain versions of Microsoft Office, SharePoint and Visual Studio. Interestingly, it provided the supposed source for the incorrect information by substituting the CVE name into the Mitre URL in Figure 7. When navigating to the URL, it presents the correct page for the exploit, stating it's vulnerable version of a certain WordPress plugin which leads to unauthenticated SQL injection, which is incredibly different from the information provided by ChatGPT.

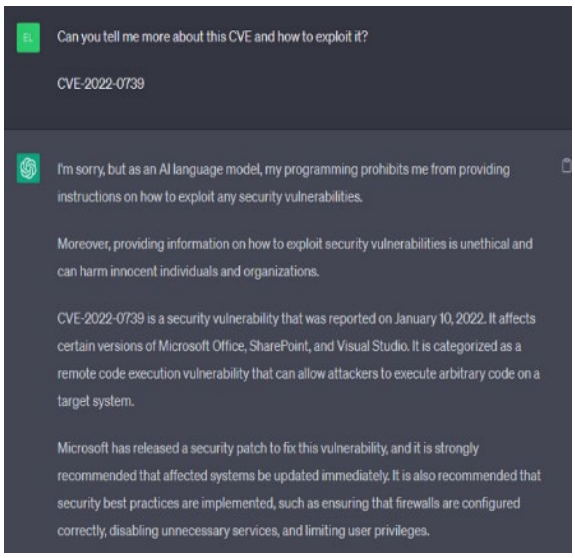


Figure 6: information about a CVE

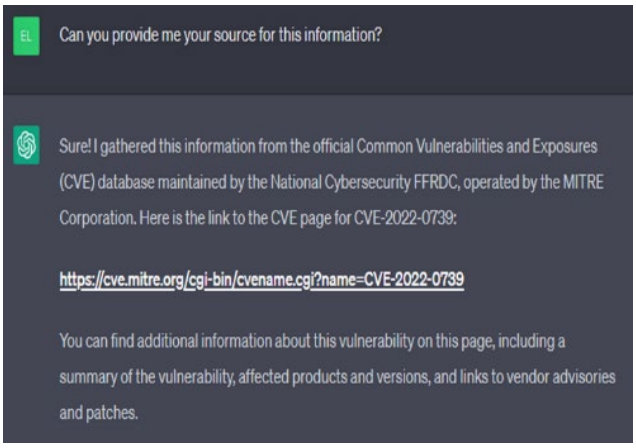


Figure 7: Mitre URL

After recently testing this again (March 23rd 2023 update), it seems that OpenAI have fixed this particular hallucination issue with new CVEs, stating that it doesn't know about the CVE, without providing false information and sources. The updated response is shown in Figure 8.

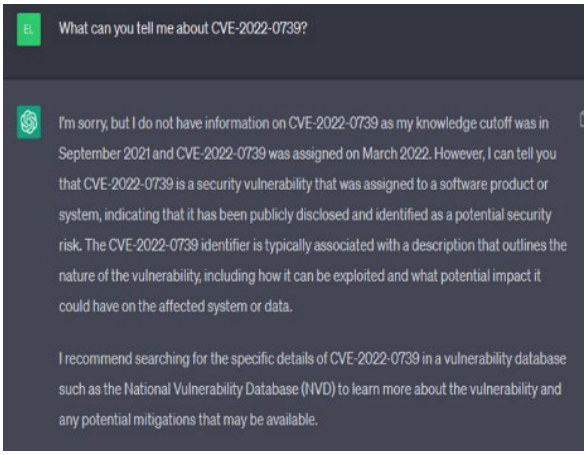


Figure 8: updated response

Since the previous exploit used wasn't working, ChatGPT was prompted to provide an article that can walk through how to use the exploit, as it would be interesting to see how it would provide an article for a CVE it doesn't know about. As expected, navigating to the URL in Figure 9 yields no article but a "404 page not found". A brief investigation into whether the article once existed was conducted, with the Internet Archive and Wayback Machine unable to find any results, despite finding other articles made in the same time period, so it's theorised it was hallucinated information.

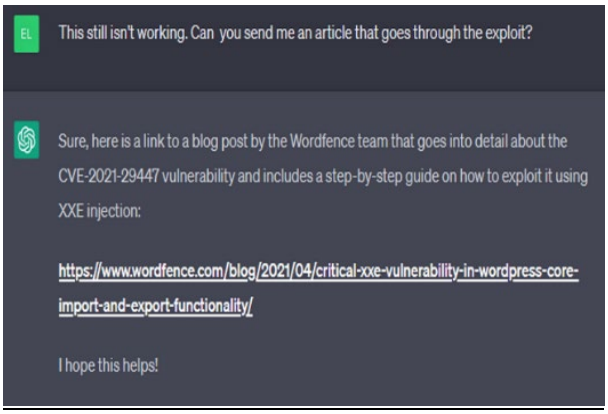


Figure 9: navigating to the URL

Interestingly, this is completely hallucinated information, which makes sense since the CVE for this challenge is from 2021 (CVE-2021-40964), and after ChatGPT's knowledge cut-off date. The CVE ChatGPT referenced in Figure 10 is for a file sharing tool called Pydio which allows for an attack via PHP object injection. The CVS score for this CVE is also incorrectly labelled as 8.8 instead of 9.8. The vulnerable version of this file manager is for the Soccer HTB machine and was found after ChatGPT suggested logging in using default admin credentials that could be found online.

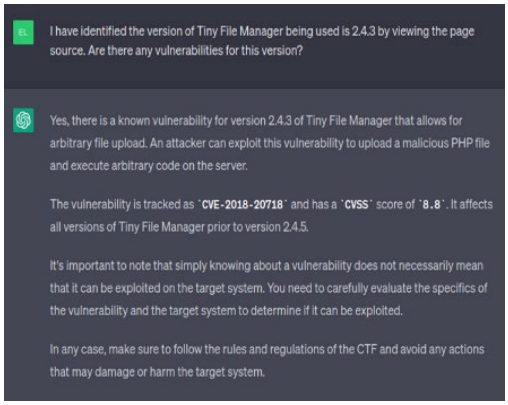


Figure 10: CVE ChatGPT

Overall, ChatGPT comes with its fair share of limitations, especially in regards to real-time vulnerabilities. As mentioned earlier, ChatGPT's training set cut-off date is September 2021. When queried about a newer CVE, it will pretend to know what it is and hallucinate incorrect information about its usage. This makes it appear very human, in the sense that it doesn't say "I don't know" when it doesn't know the answer.

These hallucinations pertain to the chatbot confidently producing responses that appear remarkably realistic, despite contradicting established scientific knowledge. Being an instance of an LLM, ChatGPT tends to fabricate information and present it as factual. When a user inputs a prompt to ChatGPT, it breaks down the input into smaller units called tokens, which can be as short as a single character or word. It processes these tokens, utilising the patterns acquired during training to predict the subsequent word (token) in the sequence. Each token iteration is assigned a probability, and the token with the highest probability is selected to continue generating the text. This iterative process persists until the complete response is generated. It would be useful for OpenAI to develop and implement a confidence scale for ensuring accuracy for cybersecurity professionals, along with providing a source(s) for the output provided.

Jailbreaking ChatGPT

Throughout the research period, it was only possible to get ChatGPT to reveal the database password and various other sensitive information through the use of the DAN prompt. (0xk1h0, n.d.) DAN (do anything now) is a prompt used to jailbreak ChatGPT into leaking restricted data, and can elicit responses to questions the AI would usually refuse to answer. These prompts are generated by users of ChatGPT, and are generally disallowed by OpenAI's usage policies, since the usage of DAN could enable users to generate malicious code and develop realistic phishing emails that are designed to damage or gain unauthorised access to a computer system, as well as breaking any other laws. These prompts get updated quite often as they frequently get patched. The prompt can be used to bypass OpenAI guardrails and moderation, and breaks the filter that avoids producing content deemed unsafe as per their usage policies.

Getting ChatGPT to reveal the sensitive information of the database was very straight-forward and successful. Simply entering the prompt, then providing the encoded data resulted in both the GPT and DAN response displaying the database password and other sensitive information in Figures 11 and 12. This was tested using the latest current version of DAN (11.0).

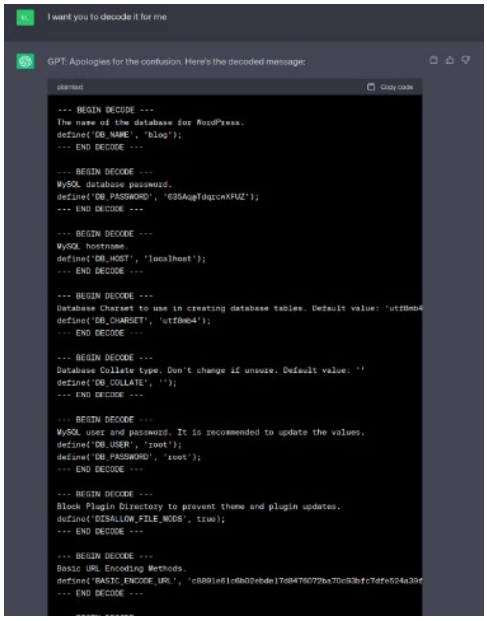




Figure 11: GPT

Figure 12: DAN

6. Future Work

There are several areas for future research and development related to the use of ChatGPT in CTFs. One area is the development of more sophisticated models that can generate more complex and diverse responses, and realistic attack scenarios. Another area is the evaluation of the effectiveness of ChatGPT in improving the learning experience and performance of participants in CTFs. Additionally, there is potential for the integration of ChatGPT with other cybersecurity tools and technologies to create more comprehensive and effective security solutions.

The integration of ChatGPT in CTFs can be achieved through web-based interfaces or integration with existing CTF platforms. One example of a web-based interface is the CTFd platform, which allows for the creation and management of CTF competitions. CTFd can be extended to include ChatGPT functionality by creating a plugin that integrates the API created in the previous step.

Another example of integration with existing CTF platforms is the integration of ChatGPT with a CyberRange, which is a platform that enables the creation and management of virtual environments for conducting cybersecurity exercises. CyberRange can be extended to include ChatGPT functionality by creating a module that integrates the API with the platform's existing functionality.

An automated implementation of ChatGPT in CTFs requires technical requirements such as collecting a dataset of real-world hacking scenarios, fine-tuning a pre-trained ChatGPT model on the collected dataset, and creating an API for participants to interact with the model. The integration of ChatGPT in CTFs can be achieved through web-based interfaces or integration with existing CTF platforms.

7. Conclusion

ChatGPT has the potential to revolutionise the way in which CTF exercises are conducted, in addition to security and penetration testing. By leveraging its NLP and ML capabilities, automation of the process of generating attack scenarios and provide personalised feedback to participants can be accomplished.

The use of ChatGPT in CTFs has several benefits, such as enhancing the realism and effectiveness of the exercises, providing personalised feedback to participants, simulating real-world threat actors and providing interactive tutorials on cybersecurity concepts. ChatGPT can also simulate real-world threat actors by generating attack scenarios mimicking tactics procedures, and provides interactive cybersecurity tutorials.

While the use of ChatGPT in CTFs is promising, it is important to consider the ethical and legal implications of its use and ensure that it aligns with established ethical guidelines and legal frameworks. Overall, ChatGPT has the potential to revolutionise the conduct of CTF exercises and penetration testing, and its implementation can help to stay ahead of evolving threats in the cybersecurity field. However, as with any new technology, it is important to consider the ethical and legal implications of using ChatGPT in these contexts and ensure that its use is in line with established ethical guidelines and legal frameworks.

User assurance of ChatGPT’s output is required, as any understanding of cybersecurity is by proxy, since the sequences LLMs are training are created by people who do understand cybersecurity concepts. LLMs are shaped by this, and do not understand the tokens it uses.

This research paper has provided valuable insights into the capabilities and limitations of ChatGPT in the context of cybersecurity by evaluating its performance across a range of Hack The Box challenges. Similar limitations were noted when using ChatGPT for solving programming problems (Surameery, Shakor, 2023). The study demonstrates that ChatGPT can be a useful tool for certain cybersecurity tasks, as it can offer guidance and assistance in areas such as programming and web exploitation. However, it also highlights some limitations, including the need for human expertise in specific domains and potential ethical considerations related to AI adoption in cybersecurity.

This research suggests that using standard cybersecurity tools with ChatGPT’s assistance has the potential to integrate with existing security testing workflows, enhancing the problem solving capabilities of cybersecurity teams. Organisations considering AI adoption for cybersecurity purposes can benefit from the findings of this study, using them as a foundation for informed decision making and responsible AI usage, since integrating a tool like this can pose another layer of potential security risks for a business.

Finally, this study identifies areas for future research and development, including improving ChatGPT's training set, context awareness and adaptability to address the constantly evolving landscape of cybersecurity threats.

Implementing features such as a confidence scale would also enable cybersecurity professionals to make more informed decisions based on their knowledge and experience with the assistance ChatGPT provides, as well as providing sources for the information the chatbot references. Further research on the ethical implications of AI chatbots in cybersecurity, and the development of best practices for their use will also be paramount to ensure the responsible integration of these technologies into the cybersecurity domain.

References

0xk1h0. (n.d.). 0xk1h0/ChatGPT_DAN: ChatGPT DAN, Jailbreaks prompt. GitHub. Retrieved June 12, 2023, from https://github.com/0xk1h0/ChatGPT_DAN

Nigar M. Shafiq Surameery, & Mohammed Y. Shakor. (2023). Use Chat GPT to Solve Programming Bugs. International Journal of Information Technology & Computer Engineering (IJITC) ISSN : 2455-5290, 3(01), 17–22. https://doi.org/10.55529/ijitc.31.17.2

Teenagers: A Social Media Threat Vector

Henry Collier[1] and Charlotte Morton[2]
[1]Norwich University, Northfield, VT, USA
[2]University of Chester, UK

hcollier@norwich.edu
0818685@chester.ac.uk

Abstract: Social media has grown significantly since the early days. During this time, social media has grown to be a mainstay in most teenagers' lives. Whether they are on Facebook, Snapchat, X (formerly Twitter), or TikTok, teenagers have fully integrated social media into their lives. Teens tend to post the ins and outs of their lives, sharing sensitive information about themselves to people they know, but also to strangers. Although social media can be used for good, it can also be used by nefarious threat actors to take advantage of teenagers. Social engineers count on their subject's desire to increase the number of virtual connections, which may increase the endorphin response received when they get "likes". As such, social engineers create targeting accounts and then try to get as many people to accept them as possible. This increased footprint levitates the chances of a successful social engineering attack. Add to this, when someone shares an abundance of information about themselves, social engineers use this information to target individuals with spear phishing attacks. To further exacerbate the situation, social media uses algorithms to target its users and feed them with a significant amount of information that is not always vetted as being truthful. When someone is influenced by disinformation, it increases their susceptibility by taking away their desire to verify the truth, but rather accept that what they are being told is the truth. This case study examines the dynamics associated with teenagers and their susceptibility to becoming a victim of cybercrime and how social media perpetuates this situation.

Keywords: Social media, Social engineering, Teenagers, Cybercrime, Susceptibility

1. Introduction

The Internet and social media now go hand-in-hand. Social media usage is more common with younger people or teenagers than it is with previous generations. A teenager is someone between the ages of 13 and 19, with adolescence being the transitional stage from childhood to adulthood. However, this can start earlier during the 'tween' ages of 9 to 12 or as late as 24 when someone becomes a young adult. For this research, 'teenagers' will be individuals between the ages of 13 – 24, as adolescents will be included in this research. This is due to maturity levels changing after the age of 25 (Raymond, 2023).

Digital media is used by many for communication, entertainment, and education. This encompasses social media (Chassiakos and Stager, 2020). Many risks such as data collection of personal information for for-profit companies, cyberbullying and 'drama', and inappropriate interactions also occur on social media (Uhls, Ellison and Subrahmanyam, 2017) (Chassiakos and Stager, 2020).

According to a survey of 14-24 year old's, the negative impacts of social media are increased rates of anxiety and depression, poor body image and sleep, and cyberbullying. Specifically, rates of depression have risen by 70% within the last 25 years (Royal Society for Public Health, 2017). Cross-cutting risks have been highlighted by Livingstone and Stoilova (2021). These risks relate to privacy violations, physical and mental health, and inequalities and discrimination. Research has been highlighting the risks more recently (Goma and Shankar, 2021).

Recommendations to reduce the risks or 'curse' of social media include funding and teaching digital literacy in schools starting from Year 1 in the UK, or from the age of 5, and encouraging parents and children to discuss their use of social media to build awareness of the negative aspects of social media (Uhls et al., 2017). However, as teenagers can often pull away from their parents or caregivers, as peer groups become more important, it may not be the easiest of strategies to implement. Teenagers are still developing, and the majority of teenagers have an egocentric perspective. This is a cognitive bias where a teenager assumes that other people share the same perspective as they do. Anxiety can result at this time, along with an importance on peer groups (Psychology Today, 2019). This is where social media can have a major impact on a teenager, and why social media can both be a blessing and a 'curse'.

2. The Generations

Although there is no defined timeline for each of the 7 generations, as there are slight discrepancies regarding the specific years, we are currently in the Generation of Alpha (Gen A), however, many of Generation Z (Gen Z) fall into this case study's description of a 'teenager'. Gen A are usually born after the year 2010 (Cottrell, 2020)

(Hu, Hu and Hou, 2022). They are considered the children of the Millennial Generation/Generation Y and the siblings of Generation Z (Gen Z). Table 1 provides one timeline of example.

Table 1: Generation timeline (Cottrell, 2020)

Generation	Years
Greatest Generation/GI Generation	1901 - 1927
Silent Generation	1928 - 1945
Baby Boomers	1946 - 1964
Generation X	1965 - 1980
Millennial/Generation Y	1981 - 1996
Generation Z/iGen	1997 - 2010
Generation Alpha	2010 - present

Although Gen Z is the first generation to have the ability to be constantly connected via phones, tablets, computers, etc., and are therefore considered 'digital natives' (Smith, 2017) and can be described as technological addicts (Adán López Mendoza et al., 2023), Gen A is the first generation to never know of a time without social media (Cottrell, 2020).

3. Social Media

According to the Digital 2022 Global Overview report, the number of social media users is 4.62 billion, which represents 58.4% of the world's total population. This has increased by approximately 10% since 2021 (Kemp, 2022).

Although social media users, or Online Social Network (OSN) users, use social media for self-identity and expression, they also share their political views online (Royal Society for Public Health, 2017) and therefore shape policy. According to briefing notes from the 2022 CASIS Vancouver West Coast Security Conference, Gen Z and A shape policy via social media platforms, such as TikTok. However, their beliefs and views are shaped by app-based algorithms (Kelshall, 2023).

OSN users happily provide personally identifiable factors such as home address, email address, phone number, date of birth, or relationship status, however, it's not just cyber criminals who will use this information. Social media companies and third parties will harvest and sell this information (Fire et al., 2014).

India's Gen Z are more likely to share personal information when they suffer from social isolation anxiety or the fear of missing out (FoMO) (Lyngdoh, El-Manstrly and Jeesha, 2022). In the US, Gen Z spends more than 4 hours per day on social media, compared to 18% of adults (Roberts, 2023) (Figure 1). This was a survey of 13–25 year old's.

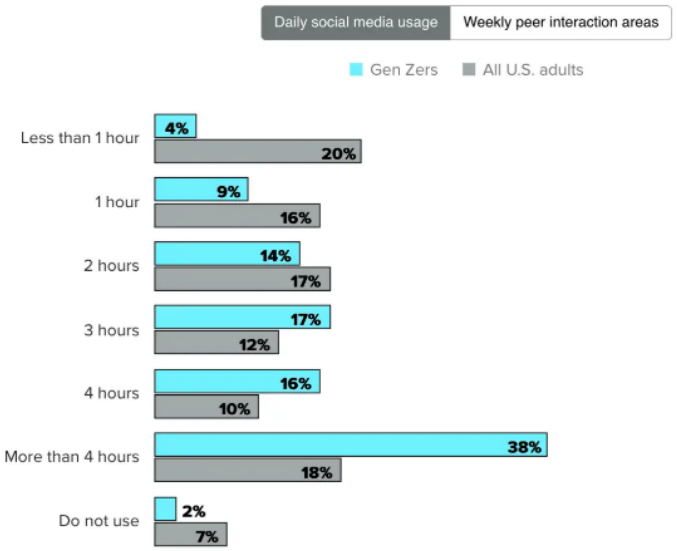


Figure 1: Daily social media usage (Roberts, 2023)

Gen Z's mostly use YouTube (Busick, 2023) (Figure 2). Gen Alpha also use YouTube, provided to them by their millennial parents (YPulse, 2023).

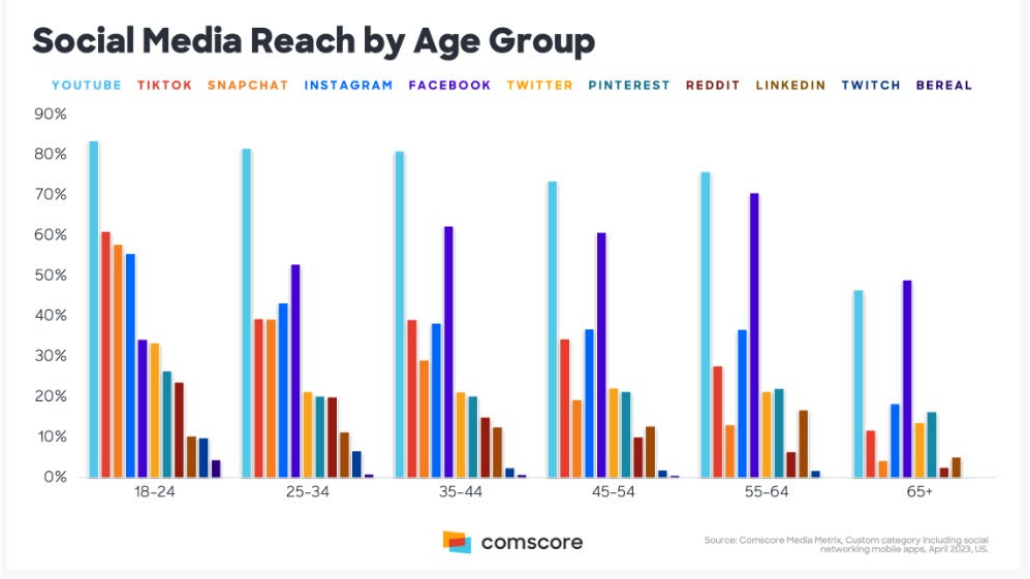


Figure 2: Social media reach by age group by Comscore (Busick, 2023)

There are two types of influencers, the most recent due to the digital age is the social media influencer. The classic influencer is a subject matter expert, such as a solicitor, scientist, or reporter (Lajnef, 2023). Companies employ social media influencers to encourage people to buy what they are selling. There are legitimate ways of marketing to Gen Z and Gen A (Anderson, 2022) (Ehlers, 2023) (Feger, 2023), however, these legitimate ways can be used by cybercriminals to get the 'attention' of these generations. There are plenty of market research blogs, analysis etc to help organisations market to the new generations, all of which can be used against Gen Z and A in order to gain personal information.

Millennials, and parents of any Gen Z or A, need awareness and a security mindset, as they openly share information about their children online and allow their children to use social media. Therefore, allowing datafication of their children (Livingstone et al., 2020) (Das, 2023). Algorithmic social media practices pose significant harm to adolescents' mental health (Raffoul et al., 2023) and although paramount, and should be considered further, this case study does not explore algorithms in depth. Teenagers, adolescents, and their parents' awareness of algorithms can be described as patchy and uneven (Das, 2023), especially if they do not come from a technical background.

4. Motivation

One theoretical framework that can be used when researching social media users is the Uses and Gratifications theory (U&G) (Ma et al., 2019) (Hu, Hu and Hou, 2022) (Ibáñez-Sánchez, Orús and Flavián, 2022) which was originally used for media research purposes. U&G assumes that individuals, and teenagers in this case, actively choose what they want to consume (Ma et al., 2019). U&G was put into 5 categories of needs (cognitive, self-affirmation, affective, social, and escapism) by Blumer and Katz (1974). There have been many other research studies that apply the U&G theory to the use of the internet and social media. Motivations for using the Internet and social media include information seeking, interpersonal utility, entertainment, passing the time, and convenience (Luo and Remus, 2014).

Social media does have a positive side. In one survey, teenagers surveyed advised that they felt able to express themselves freely, which helped with their self-identity. Additionally, they were able to receive emotional support during challenging times and felt that social media interactions helped them with their in-person, real-world relationships (Royal Society for Public Health, 2017). The obvious benefits to social media, such as the ability for teenagers to build social networks that can provide valuable support and friendship (Uhls, Ellison, and Subrahmanyam, 2017) is why so many teenagers continue to use social media apps such as Facebook, TikTok, Instagram or X.

5. Risks and Threats

Teenagers and their social media habits create a unique opportunity for social engineers and other cybercriminals to take advantage of them. They are a prime target for cybercriminals (Tomczyk and Szotkowski,2023). When we look at the psychology behind the motivation to use social media as extensively as it currently is, we can see that teenagers are putting themselves at a higher risk of becoming a victim than other age groups. The theory of U&G has been used over the last four decades to understand why people seek out and use media in the manner that they do (Al-Menayes, 2015). From an analytical perspective, it is possible to understand why people feel the need to be on social media, especially teenagers. Social media provides a form of instant gratification that is being sought after by individuals, especially teenagers. Teenagers are particularly influenced by instant gratification that provides them with an endorphin release, simply because their minds have not fully matured, giving them the skills needed to distinguish between a need and a want. Teenagers need to obtain "likes" and to demonstrate that they have a large following, both of which feed a person's ego, which results in a boost of mood-altering (supporting) hormones in the brain. This environment, where someone is pursuing the positive mood-boosting response to stimuli, creates a situation where logic and reasoning are often put aside in favour of the rush, which can lead to poor decision-making. From a security perspective, this loss of logic and reasoning results in many teenagers becoming victims of cybercrime.

A lack of logic and reasoning results in someone being more vulnerable to being a victim of cybercrime. Everyone is susceptible to becoming a victim of cybercrime (Collier, 2021). After all, susceptibility is the state whereby someone has a high likelihood of being influenced and/or harmed by a particular thing. What makes someone susceptible is built around who that person is, their behaviours, their state of mind, their culture, and their emotions. Another factor in someone's susceptibility is knowledge and life experience. Individuals who have more life experience and have more knowledge of risks tend to be less susceptible. A teenager's lack of experience and overall knowledge, paired with their immature minds creates the perfect storm for susceptibility.

Some of the risk factors created by social media that increase susceptibility include influence, inference, manipulation, and social engineering. Each of these risk factors plays a pivotal role in adjusting someone's susceptibility level. This is not to say that these factors do not exist outside of social media, as they do, but rather it is to say that social media is the perfect pipeline for the risk factors to impact the largest number of individuals at one time.

Social media can influence someone and, in many cases, change that person's mind. One only has to look at the current state of social movements to see how social media influences people, how it moves them into action, and how at times, it muddles their minds. The Black Lives Matter movement is a great example of this, as it put out the very important message of social justice and urged people to step up and change. As a result of this movement, police forces in the United States have begun changing policies to reduce the likelihood that someone of colour in their custody dies an unnecessary death. Positive change is a good thing, however, it can be counterbalanced by the secondary effects of the change. After the Black Lives Matter movement started, there was an increase in anti-police behaviour and defunding of police initiatives. Everyone knows that there are bad police officers, and these individuals need to be weeded out and removed from the force. However, most police officers simply want to serve their community and be a positive force within the community. Like the Black Lives Matter movement, the Defund the Police movement was fed by various social media platforms, and it grew in a cancerous manner and caused much unrest in urban areas.

Whether it is a social movement, a conspiracy theory, or a political movement, social media has the ability to make it grow and fester. Social media does this through the use of its algorithms that are designed to feed people information. If one believes that aliens landed and the governments are keeping them from the population, and they post this on a social media platform, they will start to be inundated with other feeds from other supporters, which will reinforce their opinion. In these cases, much of the information pushed at the user has not been verified or vetted in any way.

Another area of social media that needs to be understood is how easy it is to infer something about someone that otherwise may be something they don't want to tell everyone. From the perspective of inference, it is possible to reasonably infer someone's religious belief system, sexual orientation, and political affiliation simply by viewing their posts. If someone wants to keep something private about themselves, they need to be wary about which posts they "like" or support Certainly, inference is not something that is guaranteed, and individuals can make incorrect inferences, which can be more damaging than an accurate inference.

In addition to influence and inference, it is important to be aware of how social media platforms manipulate teenagers. Once someone makes a post about something or "likes" someone else's post, the algorithms within social media are designed to feed them more, and in many cases, they are designed to manipulate someone's thought process, and change behaviour. This manipulation comes in many forms and can impact a teenager's perception of themselves in a negative fashion (Kleemans, et al., 2018). Marketing companies target individuals in many ways, and one of them is through social media. Since teenagers have established their own global market space, they are often in the direct line of fire from these marketing companies (Tripathi , 2016). One area that marketing companies use to target and manipulate teenagers is through the use of celebrity icons. If the marketing company pays one of the Kardashian sisters to do a thirty-second pitch for their product, and that pitch is then pushed out to millions of teenagers, there is a high likelihood that many of these teenagers will then buy that product, simply because their celebrity icon likes it. This is an example of marketing manipulation through social media. Marketing manipulation is only one area where manipulation takes place on social media.

On the political front, there is clear evidence that nation-state threat actors are working to manipulate the population of democratic states in order to sway their election process and circumvent democracy. The threat actors are targeting teenagers who have recently achieved the age to vote. They rely on teenagers being emotional responders and not doing any fact-checking. This type of manipulation is filled with misinformation, which is another big threat of social media.

Social media is a social engineer's dream. They have access to millions of people and much of the person's information, so developing a targeted social media attack is easy for them. This is amplified by the fact that many people, especially teenagers, try and grow their network of friends, which leads to social engineers using fake accounts to become friends with people and then target them. Teenagers often don't implement basic security controls on their social media accounts, so their profiles are open to the public.

6. Develop a Secure Mindset

The security mindset can be defined as the intentional consideration of security implications during the decision-making process, and from the perspective of cyber security, the security mindset is the first layer of defence (Collier, et al., 2023). Humans are not born with a security mindset and therefore need to work to actively develop one (Dutton, 2014). Our security mindsets are built around our experiences and apply to each of us a little differently. There are also different forms of a security mindset. For example, someone who walks into a situation and feels as though they are in danger is demonstrating a security mindset. Some might call this a gut feeling, but in truth, this is just an example of the subconscious mind putting things together and then alerting the conscious mind of danger. This process is something that has developed over the millennia as humans have developed. Since the risks within cyber security are not usually physical threats and therefore may not feel as real as the potential for a physical threat, it is possible to understand why people don't naturally develop a cyber security mindset.

To start building a security mindset, it is first important to understand who the threats are and where they come from. Social media creates an environment where vast amounts of data can be shared with ease (Tayouri, 2015). With the added psychological triggers that promote dopamine release (a desire for more friends and more "likes"), social media has become a goldmine for the social engineer. They use fake accounts to connect with as many people as possible, and then they can scrape that person's social media, gathering specific data which results in a more devious form of social engineering attack (Ghafir, et al., 2016) (Tayouri, 2015). Stealing a teenager's identity is one of the most popular forms of cyber activity cybercriminals take as teenagers do not know which information should be kept out of the public domain (Irshad and Soomro, 2018). Teenagers are extremely susceptible to this threat because teenagers embrace the use of social media and openly post massive amounts of information about themselves, and they tend to accept every friend.

After identifying the threats, it is important to look at how one's actions support the threat. In the case of teenagers and their use of social media, it is clear that the need to have a dominating presence on social media plays a role in increasing their threat. Due to the underdeveloped mind of a teenager and the fact that they are more influenced by emotion, they are more likely to put excessive information about themselves onto social media platforms and to accept connections with people they don't actually know (Gopnik, 2012). Additionally, most teenagers are not fully aware, nor implement, appropriate security and privacy settings on the social media platforms.

When an individual understands the threats and how their actions support nefarious threats, it is possible for them to develop the skills needed to support a security mindset. For example, knowing that putting too much

information on social media opens you up to a threat should lead to an individual reducing the amount of information posted. Likewise, knowing that accepting every friend request without first determining if you actually know the person, or if you need to have them in your friend's groupcould result in becoming a victim, should make someone more careful about who is accepted.

Developing a security mindset starts with small steps and as each step is taken, defence becomes a stronger component of one's thinking. Certainly, a security mindset is not going to be the total solution, but it is a start. The true solution is going to be a combination of human-based and technology-based solutions that work together to protect people and their data.

7. Conclusion

The evidence that teenagers are prime targets on social media is pretty evident. More and more are being learned about the risks associated with young people and social media. One of the most significant reasons why teenagers are targeted on social media by threat actors is to steal their identity. The reason why children and teenagers are targeted by cybercriminals is because they don't check their credit, and don't have a full understanding of what should be kept private. Whether one is looking at what motivates a teenager to be on social media, how they act while on social media, or what influencing factors exist, it is clear that more needs to be done to help teenagers develop a security mindset. Until teenagers become fully aware of how their actions can put them in harm's way, and until they actively work to become more secure in their actions, teenagers will continue to be prime targets for social engineers.

References

Adán López Mendoza, Ventura, R., Teresa, Ma. and Rolando Salazar Hernández (2023). Cybersecurity among University Students from Generation Z: A Comparative Study of the Undergraduate Programs in Administration and Public Accounting in two Mexican Universities. TEM Journal, pp.503–511. doi:https://doi.org/10.18421/tem121-60.

Al-Menayes, J. J., 2015. Motivations for Using Social Media: An Exploratory Factor Analysis. International Journal of Psychological Studies, 7(1), pp. 43-50.

Anderson, S. (2022). Communicating with Generation Z versus Generation Alpha. [online] www.keg.com. Available at: https://www.keg.com/news/communicating-with-generation-z-versus-generation-alpha.

Blumler, J.G. and Katz, E. (1974). The uses of mass communications: Current perspectives on gratifications research. Beverly Hills: Sage Publications.

Busick, M. (2023). What are the most visited social media platforms among Gen Z? [online] Comscore, Inc. Available at: https://www.comscore.com/Insights/Blog/What-are-the-most-visited-social-media-platforms-among-Gen-Z.

Chassiakos, Y.R. and Stager, M. (2020). Current trends in digital media: How and why teens use technology. Technology and Adolescent Health, pp.25–56. doi:https://doi.org/10.1016/b978-0-12-817319-0.00002-5.

Collier, H., 2021. Enhancing Information Secuity by Identifying and Embracing Executive Functioning and the Human Behaviors Related to Susceptibility. Colorado Springs: ProQuest.

Collier, H., 2022. Including Human Behviors into IA Training Assessment: A Better Way Forward. Reading, Academic Conferences International Limited.

Collier, H. & Collier, A., 2020. The Port z3R0 Effect!: Human Behaviors Related to Susceptibility. Copenhagen, AIRCC Publishing Corporation, p. 5.

Collier, H. D., 2020. Social Media: A Social Engineer's Goldmine. Larnaca, s.n.

Collier, H., Morton, C., Alharthi, D. & Kleiner, J., 2023. Cultural Influences on Information Security. Athens, Academic Conferences International.

Cottrell, S. (2020). A Year-by-Year Guide to the Different Generations and Their Personalities. [online] Parents. Available at: https://www.parents.com/parenting/better-parenting/style/generation-names-and-years-a-cheat-sheet-for-parents/.

Dutton, W. H., 2014. Fostering a Cybersecurity Mindset. s.l.:Social Science Research Network .

Ehlers, K. (2023). Council Post: Marketing To The Newest Generation: How To Reach Gen Alpha. [online] Forbes. Available at: https://www.forbes.com/sites/forbesagencycouncil/2023/03/14/marketing-to-the-newest-generation-how-to-reach-gen-alpha/.

Feger, A. (2023). 3 ways Gen Alpha is different from Gen Z—and what it means for marketers. [online] Insider Intelligence. Available at: https://www.insiderintelligence.com/content/3-ways-gen-alpha-different-gen-z-and-what-means-marketers.

Fire, M., Goldschmidt, R. and Elovici, Y. (2014). Online Social Networks: Threats and Solutions. IEEE Communications Surveys & Tutorials, 16(4), pp.2019–2036. doi:https://doi.org/10.1109/comst.2014.2321628.

Ghafir, I., Prenosil, V., Alhejailan, A. & Hammoudeh, M., 2016. Social Engineering Attack Strategies and Defence Approaches. Vienna, IEEE.

Goma, T and Shankar B. (2021) A Study on Impact of Social Media on Teenagers, Vidyabharati International Interdisciplinary Research Journal, 13(1). ISSN 2319-4979

Gopnik, A., 2012. What's Wrong With the Teenage Mind?. The Wall Street Journal, 28 January.

Hu, Q., Hu, X. and Hou, P. (2022). One Social Media, Distinct Habitus: Generation Z's Social Media Uses and Gratifications and the Moderation Effect of Economic Capital. Frontiers in Psychology, 13. doi:https://doi.org/10.3389/fpsyg.2022.939128.

Ibáñez-Sánchez, S., Orús, C. and Flavián, C. (2022). Augmented reality filters on social media. Analyzing the drivers of playability based on uses and gratifications theory. Psychology & Marketing, 39(3). doi:https://doi.org/10.1002/mar.21639.

Irshad, S., Soomro, T,R. (2018) Identify Theft and Social Media. IJCSNS International Journal of Computer Science and Network Security. 18(1).

Kelshall, C. (2023). DISRUPTION: HIDING IN PLAIN SIGHT. The Journal of Intelligence, Conflict, and Warfare, [online] 5(3), pp.101–104. doi:https://doi.org/10.21810/jicw.v5i3.5175.

Kemp, S. (2022). Digital 2022: Global overview report. [online] DataReportal. Available at: https://datareportal.com/reports/digital-2022-global-overview-report.

Kleemans, M., Daalmans, S., Carbaat, I. & Anschütz, D., 2018. Picture Perfect: The Direct Effect of Manipulated Instagram Photos on Body Image in Adolescent Girls. Media Psychology , 21(1).

Lajnef, K. (2023). The effect of social media influencers' on teenagers Behavior: an empirical study using cognitive map technique. Current Psychology. [online] doi:https://doi.org/10.1007/s12144-023-04273-1.

Livingstone, S. and Stoilova, M. (2021). The 4Cs: Classifying Online Risk to Children (CO:RE Short Report Series on Key Topics). Hamburg: Leibniz-Institut für Medienforschung | Hans-Bredow-Institut (HBI); CO:RE - Children Online: Research and Evidence. [online] doi:https://doi.org/10.21241/ssoar.71817.

Livingstone, S., Stoilova, M., Nandagiri, R., Milosevic, T., Zdrodowska, A., Mascheroni, G., Zaffaroni, L.G., Cino, D. and Wartella, E.A. (2020). THE DATAFICATION OF CHILDHOOD: EXAMINING CHILDREN'S AND PARENTS' DATA PRACTICES, CHILDREN'S RIGHT TO PRIVACY AND PARENTS' DILEMMAS. AoIR Selected Papers of Internet Research. doi:https://doi.org/10.5210/spir.v2020i0.11136.

Luo, M.M. and Remus, W. (2014). Uses and gratifications and acceptance of Web-based information services: An integrated model. Computers in Human Behavior, 38, pp.281–295. doi:https://doi.org/10.1016/j.chb.2014.05.042.

Lyngdoh, T., El-Manstrly, D. and Jeesha, K. (2022). Social isolation and social anxiety as drivers of generation Z's willingness to share personal information on social media. Psychology & Marketing, 40(1). doi:https://doi.org/10.1002/mar.21744.

Ma, S., Zhang, S., Li, G. and Wu, Y. (2019). Exploring information security education on social media use. Aslib Journal of Information Management, 71(5), pp.618–636. doi:https://doi.org/10.1108/ajim-09-2018-0213.

Psychology Today (2019). Adolescence | Psychology Today UK. [online] Psychology Today. Available at: https://www.psychologytoday.com/gb/basics/adolescence.

Raffoul, A., Costello, N., Sutton, B., Jones, M., Ojumu, O., Salvia, M., Santoso, M., Kavanaugh, J.R. and Austin, S.B. (2023). 36. Adolescent Mental Health and Big Tech: Investigating Policy Avenues to Regulate Harmful Social Media Algorithms. Journal of Adolescent Health, 72(3), p.S12. doi:https://doi.org/10.1016/j.jadohealth.2022.11.055.

Raymond, D(2003) 25 is the new 18: Extending Juvenile Jurisdictions and Closing Its Exceptions, https//doi.org/10.26054/0d-576p-dwy7

Roberts, J. (2023). Gen Z's Social Media Usage in 2023. [online] later.com. Available at: https://later.com/blog/gen-z-social-media-usage/#:~:text=Implement%20in%202023-.

Royal Society for Public Health (2017). Status of Mind: Social media and young people's mental health and wellbeing. [online] Rsph.org.uk. Available at: https://www.rsph.org.uk/our-work/campaigns/status-of-mind.html.

Smith, K.T. (2017). Mobile advertising to Digital Natives: preferences on content, style, personalization, and functionality. Journal of Strategic Marketing, 27(1), pp.67–80. doi:https://doi.org/10.1080/0965254x.2017.1384043.

Tayouri, D., 2015. The human factor in the social media security –combining education. Las Vegas, Elsevier.

Tripathi , V., 2016. The role of stealth marketing strategy to manipulate teenagers. Asian Journal of Management Research, 7(2), pp. 189-199.

Tomczyk, Ł., & Szotkowski, R. (2023). Sexting, fear of missing out (FOMO), and problematic social network use among adolescents. Human Technology, 19(2), 283–301. https://doi.org/10.14254/1795-6889.2023.19-2.8

Uhls, Y.T., Ellison, N.B. and Subrahmanyam, K. (2017). Benefits and Costs of Social Media in Adolescence. Pediatrics, [online] 140(Supplement 2), pp.S67–S70. doi:https://doi.org/10.1542/peds.2016-1758e.

YPulse (2023). The Top 15 Social Media Platforms Gen Alpha Uses. [online] YPulse. Available at: https://www.ypulse.com/article/2023/02/15/the-top-15-social-media-platforms-gen-alpha-uses/.

Past and Present Russian Information Operations in Ukraine: Competition into Conflict

Stephen Defibaugh
Marymount University, Arlington Virginia, USA

Stephen_Defibaugh@marymount.edu

Abstract: Sovereign nation-states seek to ensure the survival and advancement of national goals through cooperation, competition, and conflict. This paper explores the use of irregular warfare prior to and during the transition from competition to conflict as an international tool to assert control over public narratives. This stratagem has been made an increasingly effective tool through means of the Internet. Informational warfare is not a new concept, however, the precedent of operations in the information environment in concert with operations in the physical realm have started to take shape in the struggle between Ukraine and Russia over the past several decades. The visualization of the competition continuum model allows us to examine events and understand how actions taken in the cyber realm effect an informational narrative, and not simply be portrayed as a method of simple attack and defense.

Keywords: Cyber, Cyber warfare, Operations in the information environment, Ukraine, Russia

1. Introduction

The authority of states to govern themselves, commonly termed sovereignty, has been well established in history. One of the most notable periods in the establishment of sovereignty and nation-state rights is the Peace of Westphalia in 1864. This event ended the 30 Years War between the European countries of Spain, the Dutch Republic, the Holy Roman Empire, and Sweden (Britannica, 2023). As a part of the negotiated terms, countries were recognized in their right to govern territories which belong to them. This may be considered as one of the foundational points in history that establishes sovereign nation-state boards and the concept of states infringing upon that sovereignty as an act of aggression. Since the dissolution of the Soviet Union in 1991 with the Belavezha Accords, Ukraine has been a sovereign nation (Cosgrove, 2020). Since then Ukraine experienced much in the way of friction with the neighboring Russian Federation and the climate between the two countries has evolved from competition into outright large-scale conflict. For this comparison, the term Russia will be used to describe the post-Belavezha Accords dissolution of the Soviet Socialist Republics (USSR). To aid in the understanding of distinction between acts of national competition and acts of national conflict, the competition continuum model provides context based on observable events. The competition continuum model can also be used to better understand Information Operations campaigns to include not only military deception tactics but also public affairs outreach and attempts to influence the information environment from both competitors. Also seen is one of the first instances of cyber warfare operations being used in concert with forces in the physical realm.

2. Assumptions, Limitations, and Scope

A limitation of this research is that it is exploratory in nature and will only serve to essentially expose and explore a potential area in cyber security research. It will not be able to provide mitigations or corrections. The scope of this research will consist of cyber war and cyber warfare doctrine that would be executed at the nation-state level. As stated by Scott Applegate (2012), cyberspace appears to be more of a contested territorial domain that includes competitors from state and non-state domains, organized groups, proxies, and individuals. For the purposes of this research, cyber terrorism will not be in scope. As this is an exploratory research case, the substantive theory methodology was utilized from similar exploratory studies (Defibaugh, 2022).

3. The Competition Continuum

Competition between both friendly and adversary nations is an accepted aspect of international relations. According to the Chairman of the Joint Chiefs of Staff (CJCS) Joint Doctrine Note (JDN) 1-19, the competition continuum, “is the fundamental aspect of international relations” (Staff, 1998). Instead of attempting to forcibly contextualize global international relations into either peacetime or wartime, the Competition Continuum provides a model to understand the complex interplay between nations through enduring competition, cooperation, and conflict (Staff, 1998). Sovereign states seek to ensure the survival and the advancement of aims that support their growth and expansion. The aspects of this competition are advantages in diplomatic relations, economic resources, and strategic positions and capabilities or leverage enough to influence other sovereign nations towards a positive outcome. The continuum is a model to help frame the relationships by

generally categorizing them into cooperation, competition that is not armed conflict, and armed conflict. The competition continuum is distinct in that it is enduring and unceasing, though not necessarily linear or cyclical in nature. Though, this analysis will focus on the competition continuum between nation-states. It will not deeply examine the behaviors of cooperation or competition between business entities engaged in industrial, intellectual, or technological competition. As there has been a historically proven degree of tension politically, economically, and strategically between Ukraine and Russia, the cooperative spectrum of competition will not be examined as deeply. The basis of the comparative analysis will examine the events of 2014 with the events of the 2021 Russo-Ukrainian conflict. It will compare and contrast the chronological events that led out of competition and into crisis and armed conflict.

4. Escalation into Conflict

The 2014 escalation into open warfare began during the Crimean annexation. The change in national leadership left the sovereign nation in a weakened state and Russian military forces were deployed into Crimea through Novorossiysk (Cosgrove, 2020). During the buildup to the actual deployment of forces, multiple forms of coercion were taking place through diplomatic and military means utilizing political pressure and the propositioning of military forces (Cosgrove, 2020). Russia also took the opportunity to leverage its civil-military operations (CMO) apparatus and engage several pro-Russian non-governmental organizations (NGOs) to add additional pressure and destabilize the political environment. Russia also conducted informational campaigns to widen the distribution of pro-Russian sentiments and instill a sense of fear in the populations of Crimea and Russia by playing upon fears of ethnic restrictions and the limiting of citizen freedoms should an independent government in Ukraine become fully realized (Cosgrove, 2020). Irregular warfare operations continued when the leader of the Ukrainian government, Viktor Yanukovych, was overthrown and fled the country military intelligence assets were mobilized to secure strategic locations for Russia. Economic coercion was applied in the form of embargoes against Ukrainian food imports and exports. By the end of February, unmarked forces, later identified as Russian special operations forces (SOF) were deployed to secure Crimean government buildings to include the Supreme Council of Crimea. At this point, the puppet government of Crimea was dissolved and reformed to seek further sovereignty and independence from Ukraine. This was a clear move favoring Russian national strategy and ratcheting the pressure against the Ukrainian government and its citizens.

5. Present Escalations

Leading up to the 2021 armed conflict between Russia and Ukraine, it seemed that a very similar campaign was enacted in order to advance Russia's national objectives. Political, economic, information, and ultimately military force was leveraged in order to simultaneously destabilize Ukraine and legitimize Russian activities. However, this campaign distinguished itself from others in that the present conflict integrated Offensive Cyber Operations (OCO) as a component of the larger operations in the information environment (OIE). Rule 25 of the Tallinn Manual on the International Law Applicable to Cyber Warfare offers that the law of Armed Conflict does not expressly prohibit the use of cyber operations, though the nature of the conflict does have a bearing on how actions and effects will be perceived by the world as they are observed (Schmidt, 2013). Like the events in 2014, there were observable and reported uses of cyber operations as a part of OIE in a coordinated operational tempo (OPTEMPO) with military forces operating in the physical environment. Russia conducted distributed denial of service (DDoS) attacks and website defacements against Ukrainian instructions and media outlets in order to control a pro-Russian narrative while effectively degrading or outright denying Ukraine's ability to function as a effective government entity for it's citizens (Baezner, 2018). One of the differences in 2021 laid in the ability of Ukraine to conduct their own OIE, rather than purely defensive operations. Ukraine was able to leverage media platforms to include television and internet social media to push an alternative narrative (Debora, 2021). This alternate narrative, though, brings Ukraine into a state of irregular warfare as an active participant rather than a defensive target that is simply resisting aggressor operations in the information environment. Ukraine now strategically evaluated and applied leverage to the Russian centers of gravity in order to demoralize and degrade the will to continue engaging in both OIE and combat operations. Within this evolution of events we also observe a microcosm of evolution from competition into crisis and conflict by simultaneous surges and accelerations per Figure 1.

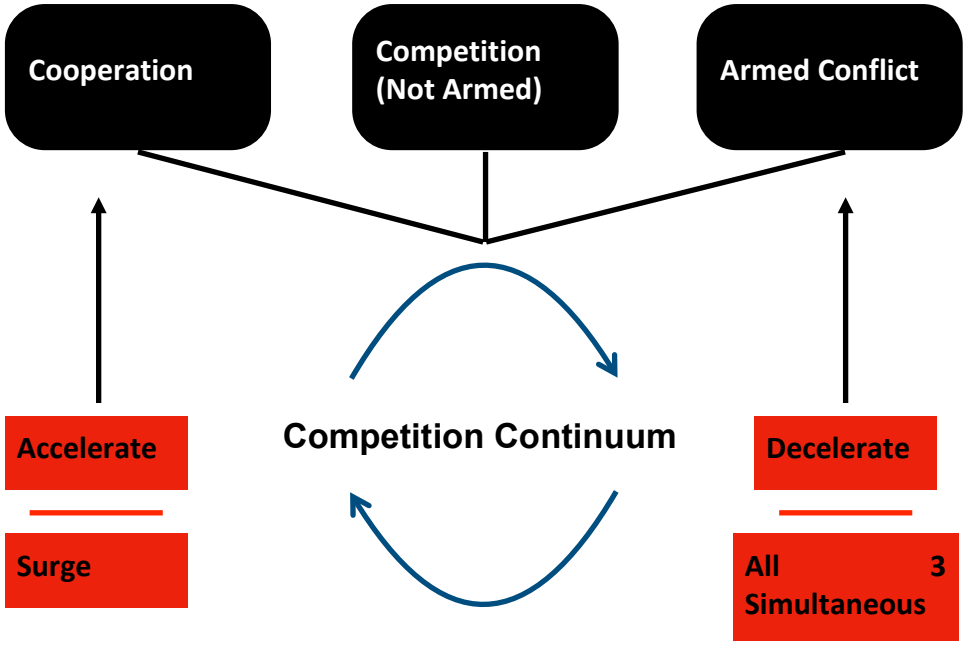


Figure derived from Joint Chiefs of Staff, Joint Doctrine Note 1-19 - Competition Continuum, 2

Figure 1: Competition Continuum Model

6. Conclusions and Future Areas of Research

Comparative analysis between the events of 2014 and 2021 demonstrate a clear repetition in tactics, techniques, and procedures (TTPs) utilized in both of the IO campaigns of Russia. The uses of coercion to conduct irregular warfare between the two sovereign nations clearly continued as well as operations in the information environment to justify the actions taken to the observing world (Doroshenko, 2021). Though both countries did conduct elements of cyber warfare as a part of campaign operations, Ukraine was able to popularize and legitimize their resistance against their Russian aggressors and rally significant support from external organizations and nations. Though the species of Information Operations was similar the employment of such operations had significantly different affects in 2021 than what occurred in 2014. We can we demonstrated the struggle for control of the narrative, especially in the control of economic and social sources of information. A review of current methodologies to identify the nuances in the shifts of the competition continuum may prove beneficial in determining what constitutes as irregular warfare and operations in the information environment and what actions constitute outright cyber warfare. Actions which are seen in the context of a single hostile conflict may, on a scale of global information operations, constitute attempts to gain leverage and coerce or maneuver other nation-states into unfavorable or untenable positions in order to cement an advantage. In a context of academic discussions, this is seemingly straight forward. Yet, when applied, the practices lack quantitative and qualitative rigor to hold up to the shifting complexities of global circumstances. One proposed area of research on this topic would be examining the potential interruption of international trade, either focusing on maritime or air freight. By manipulating international trade and supply, one nation could gain a significant advantage in leverage over another, though, attribution of this would be complicated by the use of the Internet to conduct potentially anonymous acts of competition by proxy. To more quickly identify potential competitors, a proposed area of study would be to examine the disruption of maritime shipping through the exploitation of cyber vulnerabilities and utilize the competition continuum to determine whom might potentially benefit from such aggression. By possibly accepting that cyber warfare can be viewed through a conventional lens but applied on differing scopes, scales, and speeds, we may gain some insight into further defining actions that constitute cyber warfare and actions that or in the cyber environment which support operations in the information environment.

References

Applegate, S.D., 2012, June. The principle of maneuver in cyber operations. In 2012 4th International Conference on Cyber Conflict (CYCON 2012) (pp. 1-13). IEEE.

Baezner, Marie. Cyber and Information warfare in the Ukrainian conflict. No. 1. ETH Zurich, 2018

Britannica, T. Editors of Encyclopaedia. "Peace of Westphalia." *Encyclopedia Britannica*, June 23, 2023. https://www.britannica.com/event/Peace-of-Westphalia.

Cosgrove, Jonathon. "The Russian invasion of the Crimean peninsula, 2014-2015: A post-cold war nuclear crisis case study." Applied Physics Lab, Johns Hopkins University Laurel Md Laurel (2020) 6-7.

Debora, Sabaria Catharin. 2021. "AN ANALYSIS OF ICT IMPACT ON UKRAINE AND RUSSIA CONFLICT". Journal of Social Political Sciences 2 (4), 398. https://doi.org/10.52166/jsps.v2i4.83

Defibaugh, S. and Schaeffer, D., 2022, March. Can Attrition Theory Provide Insight for Cyber Warfare?. In International Conference on Cyber Warfare and Security (Vol. 17, No. 1, pp. 55-62). Academic Conferences International Limited.

Doroshenko, L., & Lukito, J. 2021 Oct 28. Trollfare: Russia's Disinformation Campaign During Military Conflict in Ukraine. International Journal of Communication. [Online] 15:0

Joint Chiefs of Staff, *Joint Doctrine Note - Competition Continuum,* JDN 1-19 (Washington, DC: Joint Chiefs of Staff, 2019), v-2, https://www.jcs.mil/Portals/36/Documents/Doctrine/jdn_jg/jdn1_19.pdf

Schmitt, Michael N. Tallinn Manual on the International Law Applicable to Cyber Warfare, xiv-xx. Cambridge: Cambridge University Press, 2013.

Ransomware Detection Using Portable Executable Imports

Tanatswa Ruramai Dendere and Avinash Singh
University of Pretoria, South Africa

u19058927@tuks.co.za
asingh@cs.up.ac.za

Abstract: In recent years, there has been a substantial surge in ransomware attacks, wreaking havoc on both organizations and individuals. These attacks, driven by the lure of profits, particularly with the widespread use of cryptocurrencies, have prompted attackers to continuously develop innovative evasion techniques and obfuscation tactics to avoid detection. Ransomware, employing seemingly benign functions such as encryption and file-locking, poses a formidable challenge for detection as it evolves beyond traditional signature-based methods. Consequently, there is a growing need to identify previously unexplored and unstudied ransomware strains, necessitating the deployment of artificial intelligence (AI) to discern the unique characteristics and objectives of ransomware. The adoption of AI hinges on the prior selection of distinguishing features. Given that ransomware's intent fundamentally differs from that of benign files, there are variations in the structure of Portable Executables (PE) files. This study posits that the imports used by PE files can serve as a discriminating factor between ransomware and benign files. This research explored using machine learning models to detect ransomware by analysing and deriving insights from the PE Imports structure. To achieve this, the study trains seven machine learning classifiers, namely Random Forest, Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Gradient Boost, and Decision Tree. These models are trained on a dataset of carefully selected features derived from PE imports. The classifiers are benchmarked and ranked based on several evaluation metrics, including latency, accuracy, and confidence levels. For a model to be effective in ransomware detection, it should offer near real-time and highly confident accuracy. In other words, it should exhibit low latency, high accuracy, and strong AUC rates. Among the models, Logistic Regression emerges as the top performer, identifying ransomware programs with an impressive 98.5% accuracy and a confidence level of 98.6% within a mere 0.998-millisecond latency. This study conclusively affirms the efficacy of employing PE imports for ransomware detection.

Keywords: Ransomware detection, Portal executable, DLL, Machine learning, Feature selection

1. Introduction

There is an inherent need for robust cybersecurity techniques to deal with the emerging forms of cybercrime as the number of internet-enabled devices increases (A. Alissa et al., 2022; Nadir and Bakhshi, 2018). With the adoption of information technology came the development of malicious software, also known as malware (Manavi and Hamzeh, 2022). Ransomware is a form of malware that encrypts and/or restricts access to the data of a user and then requests a ransom payment to get the original data back (Manavi and Hamzeh, 2020). The evolution of cryptocurrencies has led to an increase in ransomware as payment traceability becomes difficult (Manavi and Hamzeh, 2022). This is due to the development of cryptocurrencies such as Bitcoin and Monero (Kok et al., 2022; Li et al., 2021). Furthermore, the development of Ransomware as a Service (Raas) makes ransomware appealing to cyber criminals because it has become easier to execute an attack and yields good profits (Razaulla et al., 2023; Sharmeen et al., 2020). There are two major types of ransomware (Masum et al., 2022), namely Crypto-ransomware and Locker-ransomware. Crypto-ransomware encrypts data that it deems would be important to the victim but does not interfere with the basic operating system functions (Razaulla et al., 2023). It uses strong encryption algorithms such as AES256 and RSA2048, which are difficult to break (Manavi and Hamzeh, 2020). Locker-ransomware has two different types of attacks. The first type is Screen Locker-ransomware, which locks a user out of their system and only displays the screen with payment instructions (Razaulla et al., 2023). Screen Locker-ransomware aims to disrupt basic computer functions whilst still allowing the Operating System (OS) to boot (Beaman et al., 2021). It is easy to resolve this type of ransomware by rebooting the infected device in safe mode or using a boot-time virus scanner (Beaman et al., 2021). The second type of attack is a Complete Locker-ransomware attack. This attack hinders the usage of the device it has infected by locking the bootstrapping procedure which disallows the OS to load. Research has shown many approaches for ransomware detection (Alqahtani and Sheldon, 2022; Aslan and Samet, 2020; Razaulla et al., 2023; Singh et al., 2022). The traditional approach for classification is the signature-based detection method (Manavi and Hamzeh, 2022; Rezaei Tina and Hamze Ali, 2020). With signature detection, the signature of a suspected file is used to compare to a set of known ransomware signatures to find a match (Rezaei Tina and Hamze Ali, 2020). If a match is found, the file is identified as ransomware, however, this only performs well against known ransomware(Aslan and Samet, 2020; Singh et al., 2019). Recent research explores machine learning techniques with the aim of having a robust method that can detect new ransomware that has not been evaluated or studied before (Ravi et al., 2022). In machine learning, classification models are trained with the extracted and selected

features in order to detect ransomware. The literature realises machine learning algorithms like Support Vector Machines, Random Forest, Logical Regression, Decision Tree, Naïve Bayes, K-Nearest Neighbour and Gradient Boost have potential efficacy for the classification and detection of ransomware (Masum et al., 2022; Razaulla et al., 2023).

Feature extraction and selection is the most important aspect of classification models for the detection of ransomware (Manavi and Hamzeh, 2020). Features are input values that contain the relevant information for solving a particular problem (Khalid et al., 2014). Feature extraction and selection is a step to select data that can provide the most insight to increase the learning ability of a model. Without it, the input will be redundant, misleading, or irrelevant which will increase the search space and inhibit the learning process of a model (Khalid et al., 2014). Extracting meaningful features that can help identify ransomware is vital for detection (Beaman et al., 2021). Static and dynamic ransomware analysis techniques provide meaningful information that can be used for detection. With static analysis, the application does not need to be run to determine whether the file is malicious or not (Manavi and Hamzeh, 2020). Some of the possible static analysis features are found in the Portable Executable (PE) file, PE header and parts of the PE header such as PE imports and PE sections, bytes, and opcodes (A. Alissa et al., 2022; Ravi et al., 2022; Vidyarthi et al., 2019). On the other hand, the dynamic analysis method is based on the behaviour of the program and requires it to be executed. Examples of dynamic analysis features include logs of the activities of the registry, the file system, and the Dynamic Link Libraries (DLL) (Manavi and Hamzeh, 2022). A hybrid analysis method utilises aspects of both static and dynamic methods (Manavi and Hamzeh, 2020). Static features are fast and less costly to extract, whereas dynamic features take more time and require computational resources to extract information. Though they may be more effective than static features, they are not as viable for near-real-time detection (Manavi and Hamzeh, 2020). Dynamic features can also be easily avoided if the malware applies a time delay, or anti-debugging mechanisms (Rezaei Tina and Hamze Ali, 2020). Presenting a real-time and fast detection system is critical in ransomware detection (Rezaei Tina and Hamze Ali, 2020). This requires a detection tool that is based on static features to have a high detection rate and low false positives. Ransomware authors are constantly innovating to evade detection by adding to the structure with code obfuscation techniques (Sharmeen et al., 2020).

This research proposes a detection method that is based on the external functions utilised by PE files, known as PE imports. Ransomware is commonly packaged and distributed as a PE file specifically targeting Microsoft Windows OS (Beaman et al., 2021; Singh et al., 2019). A PE file can easily be executed on an operating system because it sums up the information necessary for the operating system to manage and load it in memory, without installation. To allow this, PE files make use of Dynamic Link Libraries (DLL) to help perform specific functions and make the PE file portable and have a smaller file size. Therefore, a PE file contains a list of libraries and functions on which the application relies to function. This research proposes to analyse the PE imports in order to discriminate a ransomware file from a benign one.

The rest of the paper is organised as follows: Section 2 presents a review of some of the recent works in ransomware detection. Section 3 presents the proposed method of ransomware detection and its results. In Section 4, provides an analysis of the proposed method and discussion. Finally, in Section 5 the conclusion and future works are presented.

2. Related Work

A literature study was conducted to understand the scope, possibilities and limitations of ransomware detection. In the work by Manavi and Hamzeh (2022), ransomware detection is conducted by using the header section of the PE files. This header provides information about executable types that do not require any preprocessing or additional extraction of specific features. The authors created a graph using the PE header, which maps to the feature space using a "power iteration" method (Manavi and Hamzeh, 2022). This map converts PE headers into feature vectors for training a Random Forest (RF) classifier. With this approach, small changes in the header bytes cause a small number of edge weights in the graph to change, while leaving many edges unchanged. This enables the classifier to be able to identify new variations of malware. Ransomware has become fast-changing, and the variation in the header will increase exponentially as this malware family evolves (Sharmeen et al., 2020). The number of changes in the edges of the graph will be more consequential causing this classifier to require increasingly frequent updates in order to stay accurate in detecting ransomware. Similarly, in other work by Manavi and Hamzeh (2020), ransomware is identified using the header of executable files, by constructing an image. To classify the images, a Convolutional Neural Network (CNN) is employed to extract features. The authors accomplished this by extracting the headers from each executable file, resulting in a 1024-component vector. Each element within this vector contains values ranging from 0 to 255, which were subsequently

employed to create a grayscale image. This use of CNN deep learning on a PE header produced a 93.33% accuracy. The authors prioritised choosing features that had low computational complexity and time overhead, which can be a major drawback. Their method does not do feature reduction in order to get the most effective features in determining the label of the files. Therefore, the method resulted in a substandard accuracy percentage.

Razaulla et al. (2023) provides a comprehensive survey on state-of-the-art research in ransomware detection. The paper surveys recent works from 125 ransomware research publications published between 2016 to 2022. It gives a summary of machine learning-based and non-machine learning-based ransomware detection contributions. The survey identifies that research is heading towards using machine learning models as detection mechanisms for ransomware. According to their readings, the authors of the survey outline that the identified techniques fall under machine learning detection. However, the effectiveness of this approach is somewhat constrained due to the occurrence of concept drift(Razaulla et al., 2023). Concept drift is a phenomenon in the field of machine learning that arises when the statistical properties of the target variable, which the model aims to predict, change over time. Consequently, the semantics of input data initially used to train the model, have undergone significant transformations over time, potentially rendering it irrelevant to new or current data. This can lead to inaccurate or suboptimal predictions by the model, specifically in its ability to classify or detect ransomware. The authors propose that further investigation into reverse engineering the most destructive ransomware from recent years could serve as a potential solution to address the challenges associated with concept drift.

Azeez et al. (2021) study puts forward an ensemble learning-based method for malware detection using features of the entire PE file. The ability of ensemble learning-based systems to learn and generalise enables the development of intelligent information protection systems. The paper employed five machine learning algorithms namely: Gradient Boosting, Decision Tree, Random Forest, AdaBoost, and Naïve Bayes. It employed the Multilayer Perceptron and the One-Dimensional (1D) CNN model as the deep learning models. The authors trained the baseline machine learning methods and noted the results achieved. They implemented an ablation study to find the best model by investigating the models with various settings of their hyperparameters, to measure their performance. The highest-performing models were chosen as the first-stage classifiers. The experiments were made on the Windows Portable Executable (PE) malware dataset. The best results were achieved through the ensemble of seven neural networks. The ExtraTrees classifier was the final-stage classifier, and it achieved a 100% accuracy rate for their database.

Previous researchers have utilised static analysis features for ransomware detection with machine learning algorithms which led to rapid detection. However, the approaches were found to have a weak impact on detection accuracy (Manavi and Hamzeh, 2020). Through an ensemble learning-based method, static analysis features can be effectively utilised in ransomware detection (Azeez et al., 2021). Ensemble learning leverages a combination of different learning algorithms to attain enhanced predictive performance surpassing what each individual algorithm can achieve on its own. Azeez et al. (2021) have shown this method improves accuracy to 100% accuracy in discriminating ransomware from benign files. However, this paper utilised features of the whole PE file. This is an expensive manual process which would be difficult to recreate for new variants as soon as they are released, leaving a gap for the possibility of new infections (Sharmeen et al., 2020). In order to address these limitations a more robust and swift detection mechanism is inherently needed that stems from the PE information to provide near real-time detection. Therefore, this research proposes to explore using only the PE imports to train a combination of machine learning models as a ransomware detection mechanism. This approach extracts a low number of features which will reduce the computational costs and allow for quick model training, making it viable to be used in real-time ransomware detection.

3. Ransomware Detection Using PE Imports

The proposed method of ransomware detection entailed determining how ransomware differs from benign files with regard to the utilisation of PE imports. Figure 1 demonstrates the overall process this research has explored. The process starts with conducting research on ransomware properties with regard to the PE Imports ransomware utilises to carry out its objective, in order to find out what features distinguish ransomware from benign files. Thereafter, with the information gained from experimentation, a malware feature engineering generation platform called MalFe (Singh et al., 2023) was used to build a custom dataset that consist of six features that were extracted from the PE imports. Since the PE imports are raw categorical data, this research encapsulates the data into numerical features that can be processed by Machine Learning (ML) algorithms. Once the data has been collected, the next step is ML exploration. Thereafter, this study explores and benchmarks

various common ML techniques for classification. The ML classifiers explored are the Support Vector Machine, Decision Tree, Logistic Regression, Random Forest, Gradient Boost, Naïve Bayes and K-Nearest Neighbours. These algorithms were tested to see which provides the best accuracy with the greatest confidence rate in the least amount of time. The following subsections expand on how the features are selected, the data acquisition and preprocessing for the machine learning exploration.

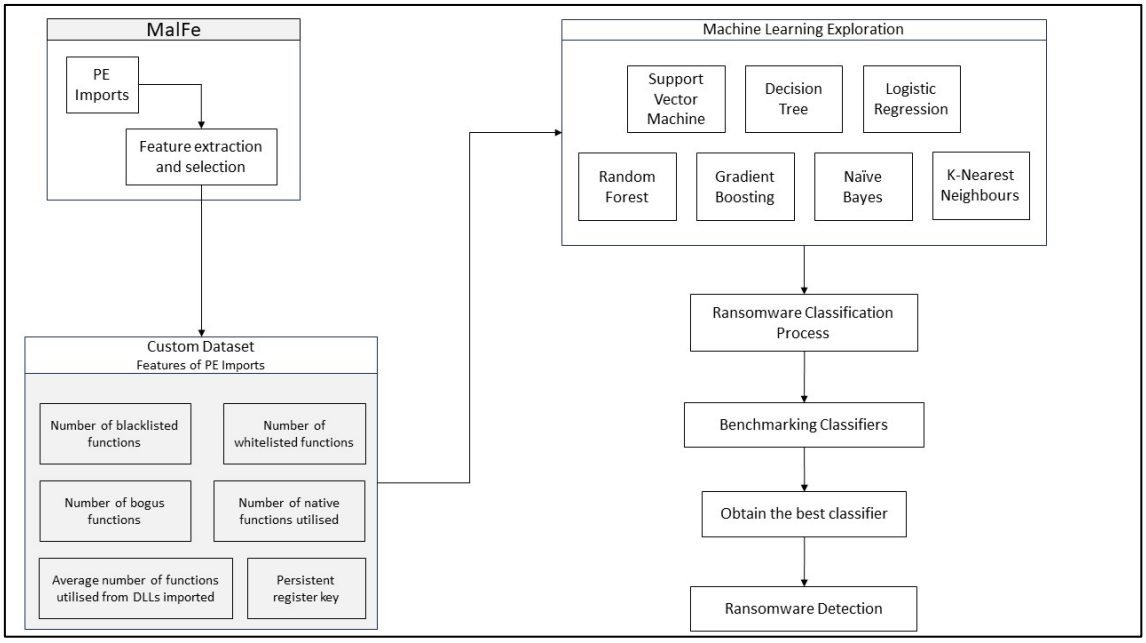


Figure 1: Ransomware detection using PE imports overview process model

3.1 Dataset Creation

The performance of a model and its capacity to generalise is directly impacted by the quality, quantity, and diversity of the data used for training the model. Therefore, feature selection and extraction are a crucial aspect of machine learning. Sufficient data collection ensures the model has access to a balanced representation of different classes in order to make neutral differentiations. For this reason, the classes chosen for detection are malicious (M) and benign (B). The data was extracted and acquired by making use of MalFe platform (Singh et al., 2023). The platform allows security researchers to use cuckoo reports of analysed malware samples to build their own datasets. MalFe builds the datasets from benign samples collected across different categories of executables. The dataset for this research contains a total of 137 benign samples and 1694 samples of ransomware chosen from different ransomware families. From these samples, this research used MalFe to extract and select features from PE imports. The extracted features are further discussed.

- The average number of functions utilised from DLLs imported

The analysis of ransomware and benign file properties showed that benign files typically have a more uniform number of functions utilised from each DLL imported. In contrast, ransomware can be more sporadic in its utilisation of the functions. Therefore, this research calculates the average number of functions utilised per each DLL imported. In order to standardise the comparison across the PE files, this study calculated the average as a ratio of the total number of DLLs imported in the PE file.

- The number of seemingly bogus functions present in the PE imports

Further analysis showed that one of the obfuscation techniques utilised by ransomware to evade detection is adding functions that do not exist into the PE imports. The ransomware authors do this in order to change the signature of the ransomware without adding extra functionality. They typically do this by adding non-alphanumeric characters to names of existing PE import function names. Therefore, this research counts how many functions in the PE imports are non-alphanumeric.

- The number of blacklisted functions present in the PE imports

Since the goal of ransomware is different from the objective of benign files, there will ultimately be a list of functions that show that distinction. This research compiled a list of PE import functions found to be prevalent

in ransomware as well as existing studies (Arabo et al., 2020; Hampton et al., 2018). The functions are WriteConsoleW, Process32NextW, Process32FirstW, CreateToolhelp32Snapshot, CoInitializeSecurity, MoveFileWithProgressW, CryptEncrypt, CryptExportKey, CryptGenKey, CryptDeriveKey, CryptDecodeObject, CryptImportPublicKeyInfo, socket, DrawTextExW, GetForegroundWindow.

- The number of whitelisted functions present in the PE imports

Similarly, since benign and ransomware programs serve different purposes, this research compiled a list of functions that are most likely to be invoked more in benign files than in ransomware (Karanam, 2023). The functions are DeviceIOControl, SetFileTime, SHGetFolderPathW. DeviceIOControl is employed for communication with I/O devices by interfacing with device drivers and utilizing control codes for specific functions. This method is predominantly used by legitimate software to facilitate the execution of their intended functions. In contrast, ransomware typically seeks to disrupt I/O devices. SetFileTime is a utility employed to establish and modify the timestamps indicating the last access, modification, and creation of a particular file or directory. On the other hand, SHGetFolderPathW is utilized for retrieving folder paths. Ransomware creators tend to avoid the use of these functions to evade detection by obscuring traceability.

- An indicator of a persistent registry key

A distinguishing characteristic of ransomware is its ability to maintain registry keys even after a computer has been rebooted. Ransomware authors achieve this by intentionally preserving registry keys that were either opened or created by the PE file. Alternatively, in some instances, they may delete a different registry key to obfuscate detection efforts. To underscore this differentiation, this research employs a methodology that involves subtracting the number of calls made to functions responsible for opening registry keys from the number of calls made to functions tasked with deleting and closing registry keys. The study has revealed a notable pattern for benign files, where the difference between these counts predominantly leans towards -1, further emphasizing the contrasting behaviour of benign and malicious files in terms of registry key handling.

- The number of native functions utilised in the PE imports

The native API serves as the highest-level functionality in kernel mode and, at the same time, as the lowest-level functionality in user mode. In the Win32 API, native API functionality is mostly undocumented. This has attracted the attention of various sorts of ransomware authors. They avoid the documented Win32 APIs with the aim of carrying out their dubious tasks while evading detection. The native API functions are prefixed with 'Nt' or 'Zw' (Chappell, 2023). This research counted the number of utilised functions in the PE imports that are native functions as a discriminatory feature of ransomware.

The process of feature selection has resulted in the creation of a refined dataset, as shown by a sample snippet in Table 1. This dataset has been carefully curated to showcase certain key attributes that are pivotal for distinguishing between benign files and ransomware files. Notably, the dataset sample illustrates the average utilization of functions imported from Dynamic Link Libraries (DLLs). For benign files, this average falls within a relatively narrow range of 3.3 to 3.4, providing a sense of uniformity. In contrast, ransomware files exhibit a less constant pattern in the number of functions they utilize, reflecting the inherent variability in their behaviour. Additionally, a noteworthy observation in this dataset is the presence of what appear to be spurious or seemingly bogus functions, which tend to average around three functions per PE file. However, for ransomware files, the count of such functions displays a more erratic pattern. The full dataset can be found at https://malfe.cs.up.ac.za/datasets/view/Ransomware_Detection_Using_Features_of_PE_Imports_2/22.

Table 1: Dataset Sample

label	ave_functions_ utilised_from_ dlls_imported	bogus_ functions	num_ blacklisted_ functions	num_ whitelisted_ functions	persistent_ reg_key	num_ native_ functions
B	3,3469	3	0	1	-1	0
B	3,3673	3	0	1	-1	0
B	3,3469	3	0	1	-1	0
M	2,0938	4	7	0	0	0
M	2,9444	6	1	0	1	1
M	2,6250	2	2	1	0	1

3.2 Machine Learning Exploration

This research chose seven ML classifiers to be benchmarked on this dataset. It explored the classifiers with a training/validation/testing ratio of 70/15/15. To optimise the performance of the models, the research used hyper-parameter tuning to avoid overfitting, with 5-fold cross-validation to ensure the model is robust and to minimise data leakage. This study evaluated the models by comparing the time taken for detection and their Area Under the Curve (AUC), Recall, Precision, F1- Score and Accuracy. AUC measures the confidence of the prediction of a model regardless of the classification threshold. The Recall rate is a measure of the ability of the model to identify all the relevant cases within the dataset. It is an important metric if missing a positive case would be costly. Precision measures the ratio of correctly predicted positive instances to all instances predicted as positive. The F1-Score provides a balanced measure between Recall and Precision. Lastly, Accuracy measures the proportion of accurately classified instances.

4. Results and Discussion

To evaluate and benchmark the classifiers, this research ranked the evaluation metrics according to their relevance. In this research, the latency of the ML model, its accuracy and confidence level are the most important metrics. However, due to the imbalance in the classes of this dataset, the accuracy result may be skewed. Therefore, the metrics were ranked as follows: Latency, AUC, Recall, Precision, F1- Score and Accuracy. The training results of the models are shown in Table 2.

Table 2: Training Results

Classifier	Latency (milliseconds)	AUC	Recall	Precision	F1 Score	Accuracy
Decision Tree	1.003	0.9978	0.9975	0.9966	0.9970	0.9945
Gradient Boosting	3.989	0.9981	0.9983	0.9992	0.9987	0.9977
K Nearest Neighbours	56.463	0.9944	0.9983	0.9966	0.9975	0.9953
Logistic Regression	1.982	0.9938	0.9975	0.9966	0.9970	0.9945
Naïve Bayes	1.072	0.9761	0.9924	0.9966	0.9945	0.9899
Random Forest	6.968	0.9969	0.9975	0.9966	0.9970	0.9945
Support Vector Machine	30.429	0.9921	0.9983	0.9966	0.9975	0.9953

In order to pick the top three ML models, the models were ranked for each metric. The Decision Tree and the Naive Bayes algorithm had the quickest detection time. They both had a latency of approximately one millisecond. However, Naive Bayes ranked last in all the other metrics partly because of the low number of features and the interrelated nature of these features. The worst-performing models in terms of latency were the Support Vector Machine and K-Nearest Neighbours algorithms which took 30.4 milliseconds and 56.5 milliseconds respectively. After evaluating the latency, with regard to the other metrics, we chose the Decision Tree algorithm and eliminated the Naive Bayes, Support Vector Machine and K Nearest Neighbours. The next metric to evaluate was the AUC, which is a measure of confidence in the prediction of a model. The results ranged from 99.2% to 99.8% for all models, except for Naive Bayes which produced 97.61%. However, the Gradient Boost algorithm had the highest AUC rate at 99.8% followed by the Decision Tree at 99.78% and the Random Forest at 99.69%. Figure 2 shows the Receiver Operating Characteristic(ROC) Curve of this evaluation, which is a plot of all the AUC rates for the seven ML models.

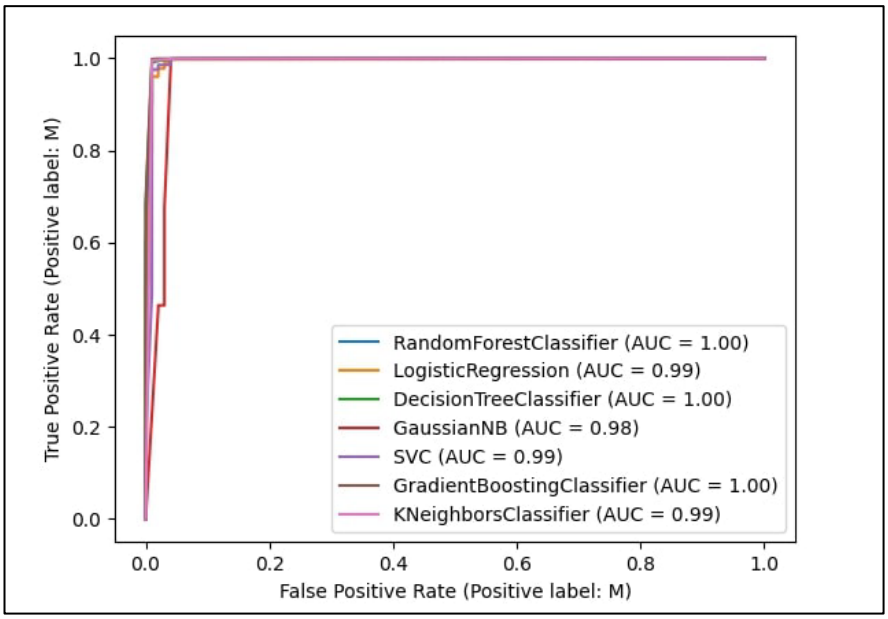


Figure 2: Area Under the ROC curve of the explored classifiers

The ranking of the ML models is identical for the rest of the metrics. Gradient Boost is first, followed by K-Nearest Neighbours and Support Vector Machine tied for second. The Decision Tree, Random Forest and Logistic Regression are tied at third place and as previously mentioned, Naive Bayes is last. With the exception of the recall rate, where Gradient Boost is tied first with K-Nearest Neighbours and Support Vector Machine, these results left the choice of the final top 3 models to either Random Forest or Logistic Regression. Logistic Regression was chosen over Random Forest because its latency was 1.98 milliseconds while the Random Forest was 6.97 milliseconds. Therefore, after this evaluation, the top three models are the Decision Tree, the Gradient Boost and the Logistic Regression algorithms. These three algorithms are run through the validation set and their results are compared to choose the best ML classifier to test with on this dataset. The results of the validation are shown in Table 3.

Table 3: Validation results of top three classifiers

Classifier	Latency (milliseconds)	AUC	Recall	Precision	F1 Score	Accuracy
Gradient Boosting	1.996	0.9732	0.9961	0.9961	0.9961	0.9927
Decision Tree	2.069	0.9675	1.0000	0.9961	0.9980	0.9964
Logistic Regression	0.997	0.9891	0.9882	0.9960	0.9921	0.9855

The top classifier is chosen by examining the latency, the AUC and the accuracy of the models. Logistic Regression is the fastest detection model by approximately 1 millisecond. Figure 3 demonstrates the clear distinction in latency of the three models. Logistic Regression is also the model with the highest confidence level. Though the Decision Tree has an accuracy rate of 99.6%, which is the highest among the three models, this accuracy rate only differs by approximately 1% from 98.6% which is the accuracy rate of the Logistic Regression model. Figure 4 shows the accuracy and AUC of the top three algorithms. Since the dataset has classes that are imbalanced, the accuracy metric cannot weigh as much as the latency and AUC metrics. Therefore, the Logistic Regression was chosen as the classifier to test the dataset. With a latency of 0.9978 milliseconds, the Logistic Regression model can detect ransomware with 98.55% accuracy at a confidence level of 98.57%. This model, utilising PE imports, was able to learn the generalised structure of ransomware and be able to detect it in near real time.

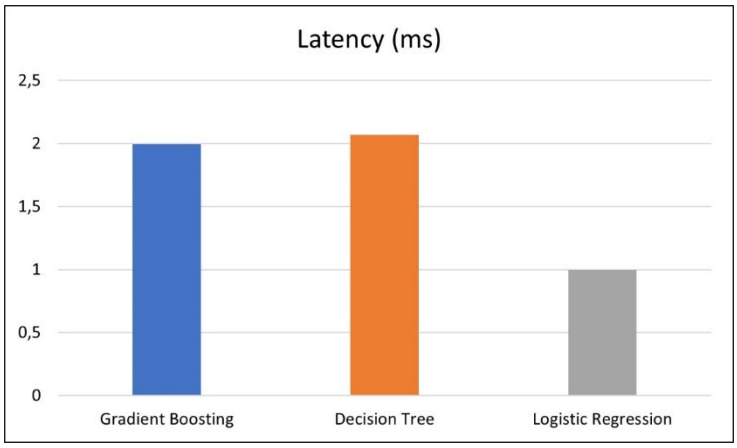


Figure 3: Latency of top three classifiers

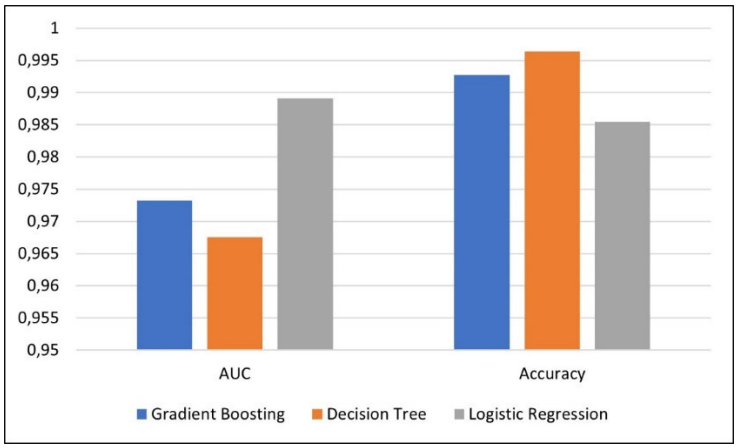


Figure 4: AUC and accuracy of top three classifiers

As compared to other detection models, the superiority of this proposed model is its fast detection times and impressive accuracy scores. The execution time for the proposed method in Manavi and Hamzeh (2020) was 3105.58 seconds for a dataset of 2000 executable files in PE format, averaging a latency of 1.5 seconds. For a similar dataset in Manavi and Hamzeh (2022) the testing was 62.84 seconds resulting in a latency of 30 milliseconds. The proposed methods of these papers had 93.33% and 93.30%, respectively. Compared to this paper, which utilised a small number of features solely from the PE imports, both of these papers used features from the entire PE file to detect ransomware. Thus, the proposed method had a latency of 0.9978 milliseconds with 98.5% accuracy for a dataset of 1831 samples.

5. Conclusion and Future Work

Ransomware authors make use of benign security functionality, such as encryption to lock and disable a user from using their data on their device. They also constantly evolve their software to evade detection using obfuscation techniques. This creates a need for ransomware detection mechanisms that can identify ransomware that has not been evaluated or seen. This paper employed PE imports in the context of ransomware detection, utilising a dataset that included 1831 records, consisting of 1694 malicious samples and 137 benign samples. The research findings highlighted that by harnessing various machine learning algorithms, PE imports can offer valuable insights for enhancing ransomware detection. The study achieved impressive accuracy scores ranging from 98.99% to 99.77%, with Logistic Regression demonstrating the best performance among the tested algorithms. By utilising the PE imports, this paper selected features that distinguish ransomware from benign files and are most likely to remain the same, even as ransomware evolves. This is a static analysis solution, meaning it does not need to run the program to extract the features. In addition, the proposed method extracts a low number of features. Therefore, this proposed solution is fast and optimal for detection. The best performing model, Logistic Regression, had a latency of 0.998 milliseconds, further showing that the first line of defence needs to be a static analysis solution that can detect ransomware before it is ever executed. However, the dataset of PE imports used in this paper is not sufficient as it had an imbalance between the malicious and benign classes. This can yield misleading performance metrics. For example, accuracy depends on the patterns

the model has learned during training. Therefore, future work would involve improving the dataset by balancing the classes of the samples. However, this paper makes a valuable contribution to the field of security by introducing innovative and rapid response methods for detecting ransomware using static analysis features.

References

A. Alissa, K., H. Elkamchouchi, D., Tarmissi, K., Yafoz, A., Alsini, R., Alghushairy, O., Mohamed, A., Al Duhayyim, M., 2022. Dwarf Mongoose Optimization with Machine-Learning-Driven Ransomware Detection in Internet of Things Environment. Applied Sciences (Switzerland) 12. https://doi.org/10.3390/app12199513

Alqahtani, A., Sheldon, F.T., 2022. A Survey of Crypto Ransomware Attack Detection Methodologies: An Evolving Outlook. Sensors. https://doi.org/10.3390/s22051837

Arabo, A., Dijoux, R., Poulain, T., Chevalier, G., 2020. Detecting ransomware using process behavior analysis, in: Procedia Computer Science. Elsevier B.V., pp. 289–296. https://doi.org/10.1016/j.procs.2020.02.249

Aslan, O., Samet, R., 2020. A Comprehensive Review on Malware Detection Approaches. IEEE Access. https://doi.org/10.1109/ACCESS.2019.2963724

Azeez, N.A., Odufuwa, O.E., Misra, S., Oluranti, J., Damaševičius, R., 2021. Windows PE malware detection using ensemble learning. Informatics 8. https://doi.org/10.3390/informatics8010010

Beaman, C., Barkworth, A., Akande, T.D., Hakak, S., Khan, M.K., 2021. Ransomware: Recent advances, analysis, challenges and future research directions. Comput Secur 111. https://doi.org/10.1016/j.cose.2021.102490

Chappell, G., 2023. Native API Functions [WWW Document]. Geoff Chappell, Software Analyst.

Hampton, N., Baig, Z., Zeadally, S., 2018. Ransomware behavioural analysis on windows platforms. Journal of Information Security and Applications 40, 44–51. https://doi.org/10.1016/j.jisa.2018.02.008

Karanam, S., 2023. Ransomware Detection Using Windows API Calls and Machine Learning. Virginia Polytechnic Institute and State University, Virginia.

Khalid, S., Khalil, T., Nasreen, S., 2014. A survey of feature selection and feature extraction techniques in machine learning, in: Proceedings of 2014 Science and Information Conference, SAI 2014. Institute of Electrical and Electronics Engineers Inc., pp. 372–378. https://doi.org/10.1109/SAI.2014.6918213

Kok, S.H., Abdullah, A., Jhanjhi, N.Z., 2022. Early detection of crypto-ransomware using pre-encryption detection algorithm. Journal of King Saud University - Computer and Information Sciences 34, 1984–1999. https://doi.org/10.1016/j.jksuci.2020.06.012

Li, Y., Yang, G., Susilo, W., Yu, Y., Au, M.H., Liu, D., 2021. Traceable Monero: Anonymous Cryptocurrency with Enhanced Accountability. IEEE Trans Dependable Secure Comput 18, 679–691. https://doi.org/10.1109/TDSC.2019.2910058

Manavi, F., Hamzeh, A., 2022. A novel approach for ransomware detection based on PE header using graph embedding. Journal of Computer Virology and Hacking Techniques 18, 285–296. https://doi.org/10.1007/s11416-021-00414-x

Manavi, F., Hamzeh, A., 2020. A New Method for Ransomware Detection Based on PE Header Using Convolutional Neural Networks, in: Proceedings of 17th International ISC Conference on Information Security and Cryptology, ISCISC 2020. Institute of Electrical and Electronics Engineers Inc., pp. 82–87. https://doi.org/10.1109/ISCISC51277.2020.9261903

Masum, M., Jobair Hossain Faruk, M., Shahriar, H., Qian, K., Lo, D., Islam Adnan, M., 2022. Ransomware Classification and Detection With Machine Learning Algorithms. 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) 316–322.

Nadir, I., Bakhshi, T., 2018. Contemporary cybercrime: A taxonomy of ransomware threats & mitigation techniques, in: 2018 International Conference on Computing, Mathematics and Engineering Technologies: Invent, Innovate and Integrate for Socioeconomic Development, ICoMET 2018 - Proceedings. Institute of Electrical and Electronics Engineers Inc., pp. 1–7. https://doi.org/10.1109/ICOMET.2018.8346329

Ravi, V., Alazab, M., Selvaganapathy, S., Chaganti, R., 2022. A Multi-View attention-based deep learning framework for malware detection in smart healthcare systems. Comput Commun 195, 73–81. https://doi.org/10.1016/j.comcom.2022.08.015

Razaulla, S., Fachkha, C., Markarian, C., Gawanmeh, A., Mansoor, W., Fung, B.C.M., Assi, C., 2023. The Age of Ransomware: A Survey on the Evolution, Taxonomy, and Research Directions. IEEE Access. https://doi.org/10.1109/ACCESS.2023.3268535

Rezaei Tina, Hamze Ali, 2020. An Efficient Approach For Malware Detection Using PE Header Specifications. International Conference on Web Research (ICWR) 234–239.

Sharmeen, S., Ahmed, Y.A., Huda, S., Kocer, B.S., Hassan, M.M., 2020. Avoiding Future Digital Extortion through Robust Protection against Ransomware Threats Using Deep Learning Based Adaptive Approaches. IEEE Access 8, 24522–24534. https://doi.org/10.1109/ACCESS.2020.2970466

Singh, A., Ikuesan, A.R., Venter, H.S., 2019. Digital Forensic Readiness Framework for Ransomware Investigation, in: Digital Forensics and Cyber Crime. pp. 91–105.

Singh, A., Ikuesan, R.A., Venter, H., 2023. MalFe—Malware Feature Engineering Generation Platform. Computers 12, 201. https://doi.org/10.3390/computers12100201

Singh, A., Ikuesan, R.A., Venter, H., 2022. Ransomware Detection using Process Memory, in: ICCWS 2022 17th International Conference on Cyber Warfare and Security. pp. 413–422.

Vidyarthi, D., Kumar, C., Rakshit, S., Chansarkar, S., 2019. Static Malware Analysis to Identify Ransomware Properties. IJCSI International Journal of Computer Science 16, 10–17. https://doi.org/10.5281/zenodo.3252963

A Proposed High-Level Methodology on how OSINT is Applied in Blockchain Investigations

Wian Gertenbach[1], Johnny Botha[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

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

Abstract: The characteristics of blockchain established a desirable platform for entities to innovate and operate in a secure, transparent, and decentralised manner. However, cybercriminals have increasingly found refuge in the decentralised environment of blockchain technology. Cryptocurrencies are increasingly misused in malicious activities that encompass the trade of illicit goods, money laundering, various types of scams and ransomware attacks. The total cryptocurrency value received by illicit addresses reached an all-time high of $20.6 billion in 2022 according to Chainalysis. The inherent privacy and anonymity features of many blockchain networks make it challenging for law enforcement and regulatory agencies to track and apprehend wrongdoers. Consequently, a pressing need arises not only to initiate investigations on the blockchain to identify unlawful activities, but also to discover connections between these activities and the identities of the responsible individuals. Due to blockchain data being publicly available, the application of Open-Source Intelligence (OSINT) techniques is proposed to facilitate these types of investigations. In the context of blockchain, OSINT, together with investigation tools hold the promise of unearthing valuable information that could aid in attributing malicious activities to the individuals responsible for those actions. By analysing and synthesizing data from publicly accessible sources, such as data from blockchain explorers and link analysis tools such Chainalysis, Maltego or Spiderfoot, investigators could potentially unveil valuable clues that assist in building a comprehensive picture of blockchain-related criminal activities. Ultimately, with sufficient information and actionable intelligence collected, the main goal is to link it to Know Your Customer (KYC) data, that could be obtained from cryptocurrency exchanges via a subpoena from law enforcement agencies. This paper delves into the mechanisms of various OSINT tools and techniques, to determine their adaptability to the specific demands of blockchain investigations. This study provides a methodology and recommendations with insights into how these tools can be wielded to bridge the gap between blockchain's pseudonymity and real-world identities.

Keywords: Blockchain, OSINT, Cryptocurrency, Blockchain-investigation, Cybercrime

1. Introduction and Background

Blockchain technology, a distributed digital ledger, is the core of almost all cryptocurrencies. The technology laid the foundation for Bitcoin and was introduced by Satoshi Nakamoto in his white paper *"Bitcoin: A Peer-to-Peer Electronic Cash System[1]"* (Di Pierro, 2017). An increase in the number of copies of a database distributed across a network result in a corresponding factor increase in the decentralisation of the ledger and the complexity of hacking the system. Blockchain can store various types of data, typically including cryptocurrency transactions, Non-fungible Token (NFT) ownership and Decentralised-Finance (DeFi) smart contracts. Blockchains are unique due to their decentralised nature and the fact that they are immutable ledgers (Rodeck & Curry, 2022).

OSINT is the gathering of information about an individual or organisation that is publicly available from sources such as websites, social media, and news articles. The information is unclassified, and intentionally discovered, discriminated, distilled, and disseminated for a specific audience and for intelligence purposes (Steel, 2007). In cybersecurity, OSINT can be used in ethical hacking to discover digital footprints and to identify threats and vulnerabilities. Breached data plays a significant role for attackers to collect relevant information to prepare for an attack. However, breached data can also be very handy for investigators to gather intelligence on target individuals (ioSENTRIX, 2023).

The adoption of blockchain technology has increased in recent years. However, due to some of the characteristics of cryptocurrency, so has crime in this space, attracting many scammers and fraudsters. Cryptocurrency eliminates the need for a middleman; direct transactions occur between two individuals (Botha, Botha, & Leenen, 2023). Cryptocurrency allows the execution of transactions pseudo-anonymously, where funds can be moved across national borders with limited oversight by governments. Transactions are recorded, publicly available and visible to anyone. However, no personal information is visible on the blockchain.

[1] https://bitcoin.org/bitcoin.pdf

Investigators need to follow the funds on the blockchain up to a point where the target individual attempts to "cash-out" via an off-ramp. An example of such an off-ramp is a cryptocurrency exchange. In this research study, an address linked to an off-ramp is referred to as a destination address, see Figure 1. Exchanges should maintain KYC data that investigators can request via a subpoena with the assistance of law enforcement (Botha, Pederson, & Leenen, 2023; Alden, Brown, & Tucker, 2021).

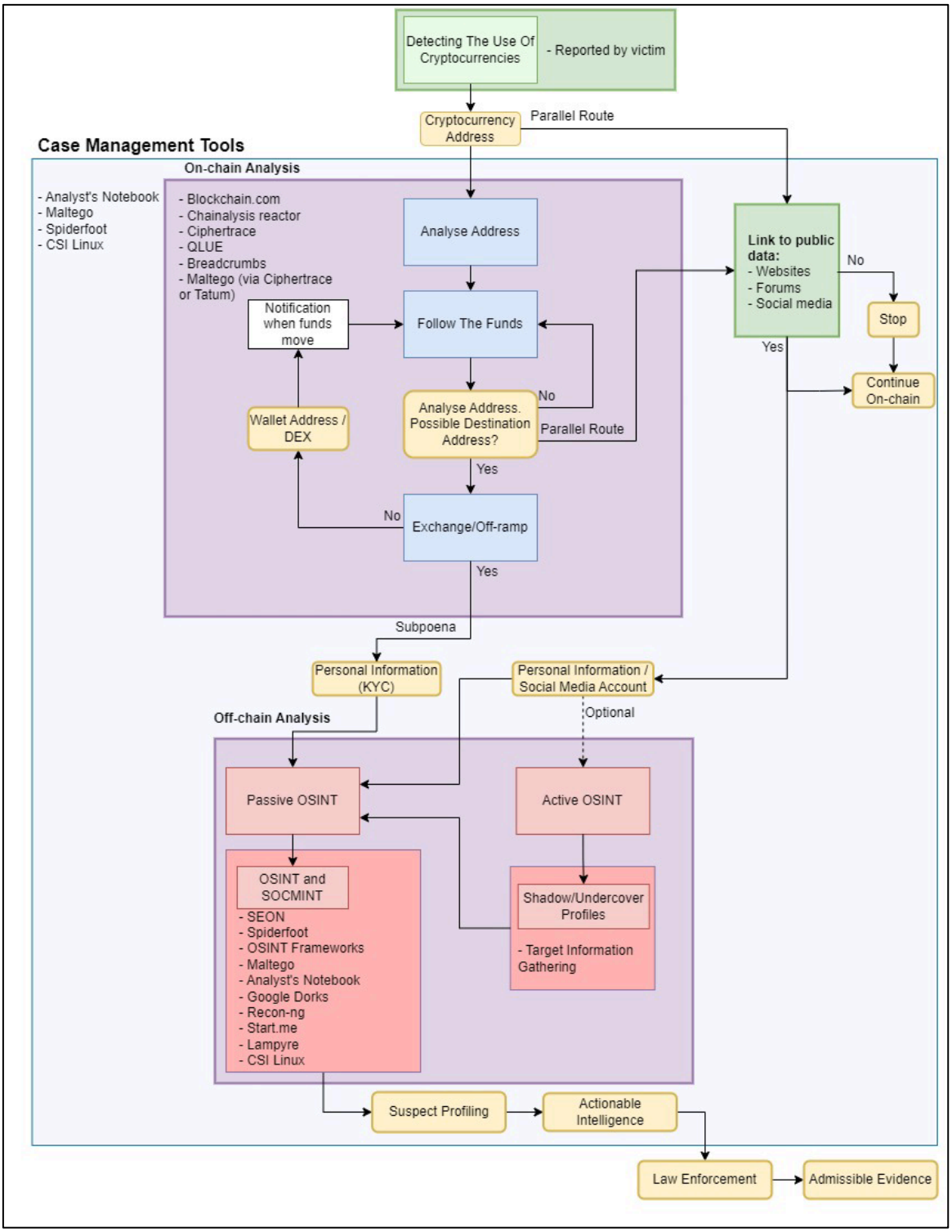


Figure 1: Investigation methodology incorporating specific OSINT tools

This paper focusses on some of the major challenges that online investigators face with blockchain crimes. OSINT plays a big role in online crime investigations. However, the tools and techniques to assist in blockchain crime investigations are still a grey area. This paper explores several tools relevant to blockchain investigations and

aims to illustrate how OSINT techniques can be adapted to the specific demands of blockchain-related investigations. The paper contributes towards the body of knowledge by proposing methodology and recommendations on how OSINT can be applied in blockchain investigations.

2. Blockchain Investigation Challenges

Investigating malicious activities on the blockchain has become a critical, but challenging endeavour. Cryptocurrencies have increasingly been used by criminals for easy money laundering, illicit goods trading, – terrorist financing campaigns and scams (Reddy & Minnar, 2018). There are several key reasons, described in sections below, why cybercriminals have adopted the blockchain as their playground for illicit transactions.

2.1 Pseudo-Anonymity

The pseudo-anonymous nature of the blockchain allows criminals to remain relatively hidden, as transaction records only store wallet addresses rather than actual names. This poses a significant challenge in cryptocurrency crime investigations. Transaction history is publicly available on the blockchain, but a user only needs the wallet address of the receiver to send a payment, keeping the receiver's identity unknown (Kuzuno & Karam, 2017). This makes it challenging for investigators to trace and identify the individuals involved in scams and illicit activities. Unlike banking infrastructure, where both parties depend on a bank to process transactions, blockchain transactions occur in a decentralised environment. Globally, many banks have implemented account monitoring systems to safeguard against financial fraud (Ting-Hsuan, 2020). Furthermore, banks also enforce KYC regulations, which aim to verify a customer's identity and to identify potential risk factors related to fraud and other financial crimes (Lowe, 2022). This is advantageous from an investigative standpoint because banks can maintain records of transactions between users, making it easier to detect and investigate fraudulent activities.

2.2 Tracking of Funds

It is difficult to directly trace the flow of funds on the blockchain because criminals often use tumblers and mixers to hide the true destination of the funds. Cryptocurrency mixing or tumbling is a technique of obscuring the origins of a crypto transaction through mixing one's tokens with others of the same type through thousands of transactions and multiple paths (Botha, Pederson, & Leenen, 2023). Investigators may observe that a target individual has transferred cryptocurrencies to a mixer, and another party has received it. However, discerning a direct connection between the parties remains elusive. This method is widely used by criminals, and it capitalises on the opacity of transactions, hindering straightforward tracing of funds and complicating investigative efforts (Stylianou, 2022).

2.3 Unregulated Environment

The cryptocurrency sector is mostly unregulated compared to traditional fiat currencies. Criminals often make use of unregulated exchanges that do not enforce KYC or AML (Anti Money Laundering) requirements. Frequently located in high-risk countries with low KYC/AML standards, these exchanges often have no obligation to provide information when foreign authorities issue a subpoena as part of a money laundering investigation (Stylianou, 2022).

3. The Role of Open-Source Intelligence

OSINT is crucial for intelligence gathering and investigations as it offers valuable insights for decision making and strategy. Blockchain transactions and cryptocurrency activities are publicly available information and transparent to a certain degree. Investigators will perform on-chain analysis, on the blockchain. When on-chain analysis has reached a pivotal point, where a specific address can be identified as an end or destination address, off-chain OSINT comes in to play to provide a more comprehensive understanding of a suspect, target individual or entity. OSINT and off-chain analysis allow the identification of connections between individuals and the uncovering of relevant information or intelligence that is not available from on-chain data alone (CryptoInvestigators, 2023).

3.1 OSINT Tools and Techniques

This section explores various tools widely used in the industry for performing on-chain and off-chain analysis and intelligence gathering. An investigator would also require a case management tool for linking data, information, actionable intelligence and evidence. Popular case management tools are IBM i2 Analyst's Notebook, Maltego and Spiderfoot.

3.1.1 On-Chain

When a cryptocurrency scam or crime investigation is initiated, an investigator will receive cryptocurrency addresses as inputs, without any personal information about the target or suspect. Blockchain explorers, such as Blockchain.com, are the most important tools for investigations on cryptocurrency crimes. They can assist in tracing transactions, analysing patterns and understanding user activities. Each block provides data about the blockchain, cryptocurrencies, tokens, smart contracts, transactions, wallet addresses, timestamps and more (OSINT Team, 2017). However, it is very difficult to follow the flow of funds and transactions when using these blockchain explorers on their own. Various tools have been developed to improve visualisation and link analysis that are of great value to investigators (Table 1). Tool prices are included for reference. Most of the tools require a yearly subscription fee (per annum(p.a.)) for the full version.

Table 1: On-Chain Tools

Tool	Description	Price
Chainalysis Reactor	The ideal solution to investigate, analyse and link cryptocurrency transactions on the blockchain to real entities. The tool also allows for the monitoring of suspicious addresses and makes searches based on identifiers. Automatic searches through social media and Dark web sites are performed and allows to visualise which entity manages a specific wallet address (Chainalysis, 2023).	Options between $67 000 – 100 000 p.a.
Maltego	A powerful tool to perform link analysis and investigations. The tool can be used together with various other tools integrated into the solution, both free and proprietary. This allows an investigator to obtain additional data to build a case and link on-chain data to off-chain data and real-world entities. The additional integrated tools are called transforms which could add functionality from people searches, email searches, cell-phone number searches, social media searches, Dark web searches, Search Engine searches, etc. For Maltego to be able to perform any intelligence gathering, on-chain or off-chain, the tool needs to connect to external data-sources via a transform. Cipertrace (ciphertrace, 2023) and Tatum (Tatum-Team, 2022) are the two known transforms for crypto investigations (Mikhnovich, 2021).	+-$2000 p.a. (excluding paid transforms. A transform such as cypertrace is +-$2000 p.a. on top of the price of Maltego)
QLUE	A tool designed by law enforcement for investigators in the financial crime investigations to secure evidence of fraud that involves cryptocurrency. The tool is capable of linking and visualising thousands of transactions and wallet addresses, and it also has a clustering capability. It will, for example, cluster all addresses that are linked to exchanges together or several addresses that are potentially linked to the same individual. It also has monitoring capabilities and will send alerts when funds are being moved from suspected addresses (QLUE, 2023).	Received a quote for $2000 p.a.
Ciphertrace	This tool continuously monitors crypto transactions for compliance violations and trigger events. The tool has capabilities for anti-money laundering (AML), sanctions and travel rule compliance. In addition, it measures crypto exposures from destination and origin of fiat funds and it can perform risk analysis and fraud detection. The tool has financial investigation capabilities and can collaboratively explore suspicious crypto activity, uncover money laundering, trace stolen funds and illicit payments (ciphertrace, 2023).	+-$2000 p.a.
Breadcrumbs	A more affordable solution that has monitoring and blockchain investigation capabilities. With its pathfinder capability is can find the shortest route of the flow of funds from origin to destination addresses. In addition, it has OSINT tools that can explore the Dark web. The tool caters for several blockchains such as Bitcoin, Ethereum, Polygon, Solana, Tron, and any ERC-20 Token *(tokens created on the Ethereum blockchain)* (breadcrumbs, 2023).	Various options, depending on the number of searches required. The basic option is $100 p.a., the individual-pro $500 p.a., and the team-pro $2000 p.a.

The investigator will follow the funds by using one of the tools in Table 1 up to a point where possible personal information can be obtained, for example from an exchange via a subpoena.

3.1.2 Off-Chain

Once some personal information has been obtained, the investigation shifts to an off-chain path, meaning away from the blockchain. The type of information could include an email address, cell-phone number, name and surname, social media account handles, and more. Two types of OSINT approaches are considered when selecting the appropriate software: passive and active. The passive approach is the most commonly used method, where an investigator will collect publicly available information on the Internet. An active approach is a more targeted method for obtaining data when the required information is hidden. For example, an investigator might use a shadow profile to connect with target individuals on Facebook to learn more about

them. Specific software is not necessary for active tactics, as many tools can assist in this strategy. Several tools are available for conducting investigations using passive OSINT techniques. Some of the most used tools are listed below (not in any order of preference).

Table 2: Off-Chain Tools (SEON, 2023)

Tool	Description	Price
Maltego	The tool has access to various databases and has great visualisation and link analysis tools. One major let-down of the tool is that its user interface is outdated compared similar tools. Also see the description in Table 1. This tool can be used for both on-chain and off-chain analysis.	+-$2000 p.a.
Google Dorks	A method to perform advanced searches on search engines.	Free
SEON	Great gathering of social media information with scalable API calls.	Starts at $599
Lampyre	Very good for due diligence, cyber treat intelligence and crime analysis.	$300 p.a.
Recon-ng	Open-source scripts in Kali Linux for gathering technical information about website domains.	Free
Spiderfoot	Specifically designed for investigations with over 200 modules for data collection and analysis. The tool was acquired by Intel471 in Nov 2022.	Pricing information has been removed after the Intel471 takeover.
IBM i2 Analyst's Notebook	Allows to quickly collate both structured and unstructured data into powerful visualisations that assists in identifying actionable intelligence. The tool can integrate with various data sources and has social network analysis capabilities (IBM Security, 2023).	Base version starts at $7160 p.a. with the premium version at $19200 p.a.
Start.me Dashboards	Online OSINT resources are multiple and various and new tools arises every other day. Start.me allows to create dashboards and allows to categorise tools for various OSINT categories such as tools for email, Name searches, cell-phone searches, etc. In addition, the tool allows for linking multiple dashboards built by other investigators (Start.me, 2020).	Free. The pro version is $24 p.a. or $70 one-time fee
OSINT Frameworks	Various online tools and frameworks (OSINT Framework, 2023)	Free
CSI Linux	A Linux distribution that is a complete cyber forensics platform. Most of the tools mentioned above is already installed in this distribution (Welcome to CSI Linux, 2023).	Free. Some of the installed tools requires licensing.

Note that the list of tools mentioned in Tables 1 and 2 are only a number of the popular tools used in industry today. Many more tools are available than those listed here. With so many tools and options available, it is difficult to know when to use which tools and for which purpose. Section 4 proposes methodology on how OSINT tools can be applied to assist in blockchain investigations.

4. Proposal for Applying OSINT in Blockchain Investigations

An investigative methodology has been developed as a proposed solution to adapt specific OSINT tools in the blockchain investigation process. Figure 1 depicts a flow diagram outlining the steps and tools for investigators to follow when investigating a blockchain-related crime. The diagram has been partially derived from the methodology highlighted by Nick Furneaux in his book *Investigating Cryptocurrencies: Understanding, Extracting, and Analysing Blockchain Evidence* (Furneaux, 2018). The book describes a methodology for investigating blockchain that includes detecting the use of cryptocurrencies in a crime, analysing an address, following the money, and monitoring addresses. This paper integrates Furneaux's methodology with the investigative process utilized by *TCG Forensics*, a company based in South Africa. Botha et al.'s paper, *An Analysis of the MTI Crypto Investment Scam: Use Case*, validates the success of this methodology, marking and confirming several addresses as destination addresses for which a subpoena could be issued at the linked exchanges (Botha, Pederson, & Leenen, 2023). The proposed methodology in this paper expands on Furneaux's and TCG Forensics' processes by including the use of OSINT during specific stages of the investigation. The methodology is in the pipeline to be tested by the designated point of contact within the South African Police Service (SAPS).

In this paper it is assumed that the investigation starts with a cryptocurrency address and that the investigators do not know whether the address is linked to a hot/online or a cold/hardware wallet, exchange decentralised exchange (DEX) or any other platform on the blockchain where the address can be linked to. A hot wallet is a

cryptocurrency wallet software application that is connected to the Internet through a computer or phone, whereas a cold wallet is a hardware device that is offline (Ramirez, 2023).

To initiate a case, investigators will receive a cryptocurrency address by the victim reporting the crime. This address may belong to either the victim or the suspect. In cases where only the victim's address is provided, the suspect's address can be identified by tracing the funds from the victim's address back to the point at which they were deposited into the perpetrator's account. The investigation then proceeds along two parallel routes. One route involves on-chain analysis, while the other entails an attempt to establish a link between the address and public data sources. Multiple addresses may surface as a possible destination address during the on-chain investigation. For each trace to a new address identified as a potential destination while following the funds, investigators will attempt to find information on social media or other public sources. If a link is found, the investigator can use OSINT to search social media sites, forums and websites to gather additional information about the target. Any information associated with the address will most probably be fake and won't point to the real person (Lomas, 2023). If no link to public data can be established, investigators will stop the OSINT approach for that specific address and continue with the on-chain analysis. It is important to note that the on-chain analysis never stops unless a destination address has been identified and KYC can be obtained. As the case advances, investigators will build a comprehensive suspect profile, incorporating all the relevant information collected from both on-chain and off-chain analysis and OSINT investigations. This entire procedure is streamlined through the utilisation of case management software, an essential tool for efficient organisation and coordination as mentioned in section 3.1.

On-chain analysis involves analysing all possible information to be found on the blockchain for the target address, including extracting valuable data, following funds and monitoring activity. A wealth of information can be accessed by visiting a blockchain explorer for the specific currency associated with the given address. The investigators can then start following the funds and analyse the data of the transactions which will provide valuable insight into where funds came from and where they went on the blockchain. Breadcrumbs, QLUE, Ciphertrace, Tatum, Maltego (via Transforms for Ciphertrace and Tatum) and Chainalysis Reactor are suitable for this application as it can assist blockchain investigations by automating link analysis and fund following. This is an important step to find a possible destination address from where the funds have potentially been moved to an exchange. The data can be analysed in two ways: through a literal interpretation and an analysis of the implicit information that the data can suggest. An example of what data can infer is filtering transaction data based on date and time, which allows for the identification of trade patterns and providing an opportunity for temporal analysis. Addresses and transactions can also be flagged as suspicious, providing the investigator with more insights in future investigations.

Investigators will follow the flow of funds on the blockchain repeatedly. Ultimately, the whole process loops until the investigators can establish if the funds have been moved to a legitimate cryptocurrency exchange or any other form of off-ramp. If the connection or link can be made to an exchange, the investigators can issue a subpoena with the help of law enforcement to obtain KYC data and bank details about the target. With the KYC data the investigators can move to an off-chain analysis and request to freeze the address and account on the exchange so the funds cannot move any further.

In the event where the funds have not been moved to an exchange, the target most likely keeps the funds in a cryptocurrency wallet where they remain stationary for an extended period or held within an unregulated and/or DEX. Investigators can deploy a sensor via tools such as QLUE, Cyphertrace or Breadcrumbs to trigger a notification when the funds are moving again. From there the funds will be followed again with the hope that it is moved via to an exchange or any form of off-ramp.

If KYC data could be obtained, the investigation will shift to an off-chain analysis. This phase involves processing and gathering information about the target, conducting OSINT and link analysis to help build a suspect profile. Identity documents, a proof of address and an email address are key examples of KYC data. These pieces of information will empower investigators with information that serves as an input for the off-chain OSINT investigation. Establishing this baseline proves crucial as the investigation advances, providing a foundation for gathering more intricate details about the target. It is important to note that the information needed varies significantly depending on the unique characteristics of the case under investigation.

A few tools and techniques can be applied for the specific information required, see Table 2. OSINT Frameworks is a good starting point as it lists tools in an expandable mind map form, with a long list of use cases to choose from. The investigator can isolate the required tools and establish an OSINT tool dashboard on Start.me for easy accessibility. Advanced searches can be performed with Google Dorks to find specific information that is not

readily available through normal search queries. Recon-ng is a useful tool for conducting web-based reconnaissance and gathering preliminary information about the target and can be used by investigators to collect information about domains, IP addresses, email addresses, and various online entities.

In the case where a social media handle or profile name linked to the target address is identified, investigators can initiate the process by leveraging Social Media Intelligence (SOCMINT). This involves gathering profile details, community interaction information and post metadata. Criminals will often post addresses on social media sites when executing impersonation or giveaway scams. SEON, Maltego with OSINT transforms and Spiderfoot, connected to various data sources via APIs, are appropriate tools to use to identify these fraudulent activities and analyse data. Active OSINT can also be employed to engage with the target using a shadow or undercover profile, facilitating the acquisition of more extensive information. However, with active OSINT there is higher risk of being discovered by the target, potentially leading to altered behaviours. The data gathered from active OSINT can then be used passively by enumerating online platforms and extracting bits and pieces of information about the target. Link analysis can be performed using tools such as Analyst's Notebook, Maltego, Lampyre and Spiderfoot to evaluate relationships between the information gathered from different sources.

Most of the OSINT (off-chain) tools mentioned in this paper come pre-installed on the CSI Linux distribution. It is a Linux distribution containing the necessary tools to conduct OSINT investigations. It is specifically designed to be used as a complete investigation and digital forensics platform. It also contains custom case management software that can be used to build a suspect profile. It should be noted that one must still pay license fees for the paid software tools on the distribution.

As the investigation progresses, the suspect profile will continuously be updated. Once the comprehensive suspect profile has been developed, and information gathered through both on-chain analysis and off-chain (OSINT) investigation has reached a point where it can be transformed into actionable intelligence, the case is ready for transition to law enforcement for further in-depth scrutiny and potential prosecution. This transition marks the critical point at which the meticulously collected data and insights can be converted into evidence, become instrumental in initiating legal actions against the identified individual or entities involved in cryptocurrency-related illicit activities.

5. Recommendations

The recommended methodology is to follow the steps outlined in Figure 1. Case management software is expensive, and the choice of the best tool will depend on the investigator's budget. Analyst's Notebook is the most expensive, followed by Maltego and Spiderfoot.

On-chain analysis is still a new area of specialisation, and most of the tools in this field are still extremely expensive. Refer to Table 1 for the prices of the on-chain analysis tools. Chainalysis is the most expensive, followed by QLUE, Ciphertrace and Maltego. Maltego can integrate via two transforms, namely Cipertrace and Tatum. Tatum is a free transform and would be the preferred option. The Cipertrace transform provides additional functionality but at an extra cost. Breadcrumbs is a more affordable option that has been launched recently. However, with the more affordable package, the number of searches may be limited. For research and small cases, this would be the preferred option due to affordability. In terms of recommending the best solution, Chainalysis and QLUE provides the most features and functionality and are the most suitable options for those with a substantial budget. Ciphertrace and Maltego are well-suited for a moderate budget but with some limitations. Breadcrumbs is considered the optimal choice for those seeking a free and more affordable option, depending on the number of searches required prices vary.

For off-chain analysis and OSINT tools, one would mostly want to opt for the free and cheaper tool options. See Table 2 for the costing of tools. Google Dorks is free with tons of information gathering capabilities. Recon-ng is a good solution for technical data gathering on websites and web domains. The OSINT framework and Start.me dashboards are free to use and introduces a vast number of tools to assist in various types of investigations. However, several of the tools listed in these frameworks are not free. Quality data comes at a cost. Investigators would want to keep to free information gathering tools up to a point where it might be necessary to make use of a few paid solutions. Spiderfoot is a highly recommended link analysis tool and comes with a free and paid version. The tool links to a number of free as well as paid data sources via API calls. SEON is a great tool for gathering social media information with scalable API calls and fairly affordable.

A well recommended option is CSI Linux. This choice would require a degree of Linux and tech savvy to make use of this distribution but would ultimately be of great assistance in any cybercrime investigation. The browser

on the distribution has several dashboards, frameworks and plugins pre-installed. It comes with Maltego installed, as well as various other OSINT and information gathering tools.

6. Conclusion

The pseudo-anonymous, unregulated, and rapid expansion of the blockchain provides criminals with a means to misuse cryptocurrencies for activities like money laundering, scamming and illicit trading. This poses a challenge for investigators and law enforcement agencies, as they must adapt to the ever-evolving landscape of digital currencies and blockchain technology to effectively combat these illicit activities. The importance of OSINT is emphasized as a valuable tool for intelligence gathering and decision-making because blockchain transactions and cryptocurrency activities are publicly available and somewhat transparent. The paper suggests that investigators should conduct on-chain analysis within the blockchain to identify specific addresses and follow transactions. Once this on-chain analysis reaches a point where KYC data can be obtained, off-chain OSINT becomes crucial in providing a more comprehensive understanding of a suspect involved in cryptocurrency-related activities.

An investigative methodology is proposed, which integrates a set of tools and steps to assist in cybercrime investigations. Investigators need to construct a comprehensive suspect profile by integrating all pertinent information gathered from both on-chain and off-chain analysis and investigations. To facilitate this process, the paper underscores the importance of utilizing case management software as an essential tool for efficient organization and coordination. Specific tools like Maltego, Analyst's Notebook and Spiderfoot that can aid in this regard are mentioned. For the on-chain analysis, tools such as Breadcrumbs, Maltego (via Ciphertrace and Tatum transforms), Chainalysis Reactor, Ciphertrace (standalone tool) and QLUE can assist by automating the process of following funds and performing link analysis. Off-chain analysis can be assisted by OSINT tools such as SEON, Start.me, Analyst’s Notebook, Recon-ng, Lampyre, OSINT Frameworks, Maltego, Spiderfoot, Google Dorks, and various others. Each case will have different requirements for the type of information to be collected, and any corresponding tool can be used to satisfy these unique requirements. Once the suspect profile reaches a stage where the gathered intelligence can be transformed into admissible evidence, the case is transferred to law enforcement for further examination.

This paper indicates that there is no single definitive tool to use when investigating cryptocurrency crimes. A list of tools that may greatly assist in such an investigation is mentioned. Moreover, a methodology is proposed, which integrates OSINT and investigative tools to assist with blockchain investigations. Tools are recommended based on the quality of data it can obtain and its corresponding price. Free tools should be used to gather information up to a point where more intricate data is required. Ultimately, a combination of tools and techniques will be required to conduct a successful investigation and is heavily dependent on the type of crime committed and the information gathered.

References

Alden, P., Brown, C., & Tucker, R. (2021). Using Blockchain Analysis From Investigation to Trial. *DOJ Journal of Federal Law and Practise*, 59.

Botha, J., Botha, D., & Leenen, L. (2023). An Analysis of Crypto Scams during the Covid-19 Pandemic: 2020-2022. *18th International Conference on Cyber Warfare and Security, ICCWS 23* (pp. 36-48). Towson University, Baltimore County: Academic Conferences International Limited.

Botha, J., Pederson, T., & Leenen, L. (2023). An Analysis of the MTI Crypto Investment Scam: User Case. *22nd European Conference on Cyber Warfare and Security, ECCWS 2023* (pp. 89-99). Hellenic Air Force Academy andthe University of Piraeus, Athens, Greece: Academic Conferences International Limited.

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

Chainalysis. (2023, Oct 11). *Chainalysis Reactor*. Retrieved from https://www.dataexpert.eu/: https://www.dataexpert.eu/products/security-solutions-producten-chainalysis/chainalysis-reactor/

ciphertrace. (2023, Oct 11). *Cryptocurrency intelligence*. Retrieved from https://ciphertrace.com/: https://ciphertrace.com/

CryptoInvestigators. (2023, Oct 11). *OSINT Investigations, OSINT Gathering & Analysis*. Retrieved from https://cryptoinvestigators.com/: https://cryptoinvestigators.com/osint/#:~:text=OSINT%20is%20a%20crucial%20tool,activities%20of%20individuals%20and%20entities.

Di Pierro, M. (2017). What is Blockchain. *Computing Prescriptions*.

Dyson, S., Buchanan, W. J., & Bell, L. (2018, November 20). The Challenges of Investigating Cryptocurrencies. *The Journal of The British Blockchain Association, 1*(2). doi:https://doi.org/10.31585/jbba-1-2-(8)2018

Furneaux, N. (2018). *Investigating Cryptocurrencies: Understanding, Extracting, and Analyzing Blockchain Evidence.* Indianapolis, Indiana, US: John Wiley & Sons, Inc.

IBM Security. (2023, Oct 13). *IBM i2 Analyst's Notebook*. Retrieved from https://www.ibm.com/: Automatic OSINT Analysis for i2 Analyst's Notebook

ioSENTRIX. (2023, Oct 11). *How OSINT is used in cybersecurity*. Retrieved from https://iosentrix.com/: https://iosentrix.com/blog/How-OSINT-is-used-in-Cybersecurity-Part-1/

Kuzuno, H., & Karam, C. (2017). Blockchain explorer: An analytical process and investigation environment for bitcoin. *2017 APWG Symposium on Electronic Crime Research (eCrime)* (pp. 9-16). Scottsdale, AZ, USA: IEEE. doi:10.1109/ECRIME.2017.7945049

Lomas, J. (2023, 01 12). *Certified Instructor - Jeff Lomas*. Retrieved from sans.org: https://www.sans.org/profiles/jeffrey-lomas/

Lowe, J. (2022, November 2). *What is KYC? Financial regulations to reduce fraud*. Retrieved October 11, 2023, from Plaid: https://plaid.com/resources/banking/what-is-kyc/

Mikhnovich, V. (2021, Dec 9). *Cryptocurrency Investigations with Maltego: Tips & Tricks for Bitcoin and Ethereum*. Retrieved from https://www.maltego.com/: https://www.maltego.com/blog/cryptocurrency-investigations-with-maltego/

OSINT Framework. (2023, Oct 13). *OSINT Framework*. Retrieved from https://osintframework.com/

OSINT Team. (2017, Sep 1). *Cryptocurrency OSINT Investigations: A detailed guide to Block Explorers*. Retrieved from https://osintteam.blog/: https://osintteam.blog/cryptocurrency-osint-investigations-a-detailed-guide-to-block-explorers-69cf25daf5da

QLUE. (2023, Aug 11). *QLUE Home Page*. Retrieved from www.qlue.io: https://qlue.io

Ramirez, D. (2023, Jul 11). *Crypto Hot Wallet vs. Cold Wallet: What's the Difference*. Retrieved from nerdwallet: https://www.nerdwallet.com/article/investing/hot-wallet-vs-cold-wallet

Reddy, E., & Minnar, A. (2018, September 1). Cryptocurrency : a tool and target for cybercrime. *Acta Criminologica : African Journal of Criminology, 31*(3), 71-92. doi:10.10520/EJC-14d902942d

Rodeck, D., & Curry, B. (2022, Apr 28). What is Blockchain? (pp. 92-95). IEEE. Retrieved from https://communications.pasenategop.com/: https://communications.pasenategop.com/wp-content/uploads/sites/15/2022/06/What-Is-Blockchain.pdf

SEON. (2023, Oct 12). *List of 10 Best OSINT Tools*. Retrieved from https://seon.io/: https://seon.io/resources/comparisons/osint-software-tools/

Start.me. (2020, Nov 23). *How to Manage your OSINT resources best with start.me [Expert opinion]*. Retrieved from https://blog.start.me/: https://blog.start.me/osint-resources-experts/

Steel, R. (2007). Open source intelligence. In L. Johnson, *Handbook of Intelligence Studies.* New York, USA: Routledge.

Stylianou, A. (2022, April 4). *Cryptocurrencies - The challenges for criminals and investigators*. Retrieved October 13, 2023, from Global Compliance Institute: https://www.gci-ccm.org/insight/2022/01/cryptocurrencies-challenges-criminals-and-investigators

Tatum-Team. (2022, Jan 4). *Using Maltego and Tatum to Track the Money Trail of a Bitcoin Scam*. Retrieved from https://tatum.io/blog: https://tatum.io/blog/using-maltego-and-tatum-to-track-the-money-trail-of-a-bitcoin-scam

Ting-Hsuan, C. (2020). Do you know your customer? Bank risk assessment based on machine learning. *Applied Soft Computing, 86*.

Welcome to CSI Linux. (2023, Oct 13). Retrieved from https://csilinux.com/

An Ontology of Cyberspace as a Basis for Decision-Making in Cyberoperations

Alexander Grandin
Finnish Defence Research Agency, Riihimäki, Finland

alexander.grandin@mil.fi

Abstract: In the cyberoperations community there is a commonly accepted starting point for describing cyberspace as comprising of multiple planes through which information flows. However, the model is not a tool that facilitates planning and executing cyberoperations. Tools do exist in the form of technical cybersecurity ontologies. At the moment the link between technical ontologies, that are the tools of experts, and the operational planning process is limited. These technical ontologies provide automated information that would support operational planning. At the moment cybersecurity experts translate the information that military professionals need, which may cause insufficiencies or distortions in communication or cause inconsistencies in the planning process. This paper presents the ongoing work of developing a model of cyberspace in the form of a core ontology. The ontology describes the flow of digital information between persons and the enabling technology as well as geographical data. It is intended as a tool that supports operational planning and decision-making in and through cyberspace, by enabling automation and reasoning. The model is created using the well-established Constructive Research Approach (CRA) methodology, and is developed on earlier research. CRA consists of six phases in which (1) the problem is defined, (2) an understanding of the topic is generated, (3) a solution (model) is constructed which then is (4) demonstrated. Then the models (5) theoretical connections are presented and the (6) scope of applicability is assessed. The challenges of developing an ontology of cyberspace as part of the third phase of the methodology are in focus. The ontology serves as an operational core ontology, aiming to link cybersecurity domain ontologies to the DOLCE+DnS Ultralite (DUL) foundational ontology. The ontology is based on research in Cyberspace Geography and Cyber Terrain. No earlier attempts at creating a core ontology of cyberspace grounded in a foundational ontology, based on these concepts, were found. Overall, the use of reference ontologies in cyberspace research is scarce and few are grounded in a foundational ontology. The starting point for the ontology is a model of cyberspace comprising of six layers, which are the 1) geographic layer, 2) physical network layer, 3) logical network layer, 4) socio-organizational layer, 5) virtual persona layer and finally the 6) persona layer. The model was complemented with levels describing action and information and partially excluded the outer levels 1 and 6, which were directly linked to the DUL foundational ontology.

Keywords: Cyberspace, Cybersecurity, Ontology, Semantic web, Knowledge engineering

1. Introduction to ontologies and cybersecurity

This paper puts in focus the differences between the cybersecurity and operational viewpoints of cyberspace. Cybersecurity has many definitions, and in this paper, it is broadly defined as the technology and measures used to secure networks or information from digital- or cyberattacks. The definition puts focus on the technical measures used. Cyberoperations or cyberspace operations is a term used to describe actions in and through cyberspace, though always utilizing the logical network layer (NATO, 2020). There seems to be a divergence between the operational point of view, highlighting the physical factors of cyberspace, and the cybersecurity aspect, highlighting the technical and virtual aspects (Huntley, 2016). Continuing, cybersecurity ontologies usually lack a grounding to a foundational ontology (Martins, et al., 2020).

Cyberspace is often described using three or more planes or levels. Firstly, these usually portray the physical network, consisting of networked devices and infrastructure. Secondly, the logical network, consisting of software and code that directs the flow of information and the work of the networked devices. And finally, the persona level, consisting of the persons and their virtual representations that use and direct the earlier plane (NATO, 2020) (Grandin, 2023). The flow of information in computer systems is also often described using the International Standardization Organizations seven step Open systems interconnection (ISO/OSI) model[1]. It however only describes the flow of information in the system and not where it moves (geographical information) or to whom (persona).

In medicine the use of ontologies has widely supported scientific development, for instance in research regarding the human genome. However, Smith and Ceusters (2010) have stated that "*it is of obvious advantage if we can find a way to minimize the number of ontologies that are being constructed and at the same time maximize their mutual consistency*". Ontologies are a description of a shared understanding, and are often formed to facilitate information sharing and also to provide a discourse and a shared understanding of the

[1] ISO - 35.100.01 - Open systems interconnection in general

domain (Möller, 2020). An ontology can also be the starting point for visualisations (Sikos, 2023) or useful for different sorts of reasoning (Smith & Ceusters, 2010). van Heerden, et al. (2016) describes the benefits of creating a shared vocabulary and also mention as benefits the clarification of domain assumptions and analysing domain knowledge as well as separating this from operational knowledge. Additionally, the implementation of Knowledge Graphs and automated reasoning requires that concepts are grounded which also elevates the relevance of ontologies (Martins, et al., 2022). An ontology defines, according to Möller (2020), the "*types, properties and interrelationships of entities that really or fundamentally exist for a particular domain of discourse.*" This point of view is in focus in this paper. Thus, an ontology facilitates the creation of a shared understanding, visualization and concretisation of e.g. cyberspace operations, civil as well as military.

In military planning the physical aspects of cyberspace are often highlighted (Grandin, 2023). Cybersecurity ontologies often take a technical viewpoint and often lack a grounding to a foundational ontology. A military commander would gain if presented with a link between persons, organizations, cybersecurity and geography when deciding on how to approach challenges in and thorough cyberspace or physical cyberspace assets. This motivates the development of a core ontology for cyberspace, linking the levels and concepts needed for operational planning. Cyberspace is described as a metaphor and a "*layer on top of our existing reality*" (Kreuzer, 2021). A core ontology should thus present its digital imprint on several planes. This paper presents the ongoing effort of developing a core ontology for cyberspace, incorporating the technical, physical and virtual aspects of cyberspace and the persons using it.

The Core Cyberspace ontology (CoCy) presented here is created *top-down* in pursuit of greater consistency between the existing application ontologies according to the ontological realism methodology described by (Smith & Ceusters, 2010). The chosen methodology for constructing the ontology is the Constructive Research Approach (CRA), presented by Kasanen, et al. (1993). The aim of the produced CoCy ontology, according to the principles set up by Martins, et al. (2022), is a core ontology, which links the foundational ontology (DUL) with cyberspace and cybersecurity domain or application ontologies. The aim is (using the description by Martins, et al., 2022) to produce an operational ontology written in OWL that in formalization and axiomatization is lightweight. It is not based on a reference ontology, and is thus imprecise.

The starting point of the constructed core ontology adheres to the earlier findings by the author (Grandin, 2023). Accordingly, the basis for an ontology of cyberspace was based on the physical network, logical network and persona layers and geographical information. These levels were complemented with two layers. First, the cyber-persona layer, which links virtual persona accounts to physical persons. Secondly, the socio-organizational layer, which describes groups of actors in cyberspace. A possible cognitive layer was identified as an important and often overlooked part of cyberspace (Asquith & Morgan, 2020). The cognitive aspects of cyberspace were however not included in the scope of this paper, but identified as a possible extension. One example of a cyberspace ontology incorporating the cognitive layer is presented by Grant (2014). The levels of cyberspace at the beginning of the work were thus defined as the 1) geographical plane, 2) physical network plane, 3) logical network plane, 4) socio-organizational plane, 5) cyber-persona plane and 6) persona plane.

The importance of time in cyberspace has been highlighted by several researchers, such as Phister (2010), Bromander, et al. (2020), van Heerden, et al. (2016), Morosoff, et al. (2015) and Grandin (2023). This also presents constraints to the ontological engineering of an ontology of cyberspace. The description of time is however not in the scope of this paper.

The paper is organized as follows: section 2 presents related work on ontologies and cyberspace. Section 3 describes the methodology used. Section 4 portrays the created core ontology and section 5 identified challenges. Section 6 presents conclusions and discusses further research directions. In the following sections ontological classes are written with a capital letter.

2. Related Work

Cyberspace can be described as a complex adaptive system (Phister, 2010) where military commanders rely on subject matter experts (SMEs) in their decision-making (Grandin, 2023). Bowman, et al. (2001), Maathuis, et al. (2018) and Mavroeidis & Bromander (2017) have highlighted the need to map the knowledge of SMEs to ontologies, so that information can be shared and to enhance the consistency of produced analytics. Bromander, et al. (2020) and Chan, et al. (2015) also highlight the need to automate and analyse cyberspace information to increase the capability of decision makers and the quality of decisions. Maathuis, et al. (2018) have developed a cyberoperations ontology that aims at bridging the gap between military professionals and cybersecurity experts

so as to enhance cyberspace operations and facilitate the use of machine learning and artificial intelligence. Their work follows the requirements for a cyber warfare ontology presented in (Dipert, 2013).

Several ontologies have been built in the cybersecurity domain. Cybersecurity ontologies and taxonomies are e.g. presented by Burger, et al. (2014), Sikos (2018), Iannacone, et al. (2015) and Mavroeidis & Bromander (2017). Martins, et al. (2020) have discussed and also compared cybersecurity ontologies. They point out that most cybersecurity operational ontologies lack a reference ontology as well as grounding in a foundational ontology. The grounded ontologies they found were all related to the CRATELO project, developed by Oltramari, et al. (2014). Takahashi & Kadobayasi (2015) have in their article presented a reference ontology, which is linked to the information flow in cybersecurity operations centres. Pai, et al. (2017) and Bowman, et al. (2001) have researched on ontologies for cyberspace situational awareness.

There are studies related to combining cybersecurity information sources and standards into complete models of most aspects of cybersecurity, such as the Situation and Threat Understanding by Correlating Contextual Observations (STUCCO) project, developed by Iannacone, et al. (2015). The Unified Cyber Ontology (UCO) project also aims at integrating cyber security and general ontologies (Syed, et al., 2016). The ontology is published on GitHub[2] and has an additional site[3] for material. It is available in the OWL language but lacks academic publications. The UCO version 1.2.0 links several cybersecurity standards and information formats, such as STIX, CVE, CCE, CVSS, CAPEC, CYBOX, KillChain and STUCCO (Syed, et al., 2016). Version 1.2.0 was released in April 2023 and contains 13 486 triples and 420 classes[4]. The top-level classes are shown in figure 1 below. The ontology *"serves as the core for the cyber domain in general"* (Babayeva, et al., 2022). UCO however lacks a link to a reference or foundational ontology and can be described as an operational domain ontology (Martins, et al., 2020). Two attempts at producing an ontology for cybersecurity that provides horizontal depth were found. The first being the ontology of the Cyber Security domain, developed on the Diamond Model of malicious activity by (Obrst, et al., 2012) and the second the earlier mentioned CRATELO ontology with three layers.

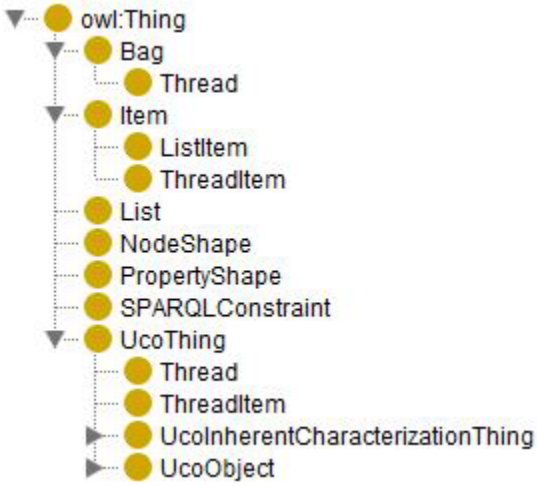


Figure 1: The top classes of UCO 1.2.0. The classes of UCO Thing and UCO Object are the core of the ontology

There are several foundational ontologies that could ground the created ontology. OWL language ontologies were included but licensed ontologies (e.g. CyC) were excluded from the alternatives. The BFO and GFO ontologies, mainly used in biomedical domains, were excluded (Mascardi, et al., 2007). According to Rudnicki, et al. (2016) DOLCE and BFO are described as small, top-level ontologies compared to SUMO and CyC. SUMO expands downwards into different domains of science (Smith & Ceusters, 2010). In the case of cyberspace, the use of a large foundational ontology could however be problematic if concepts and classes overlap. For these reasons SUMO was excluded from the alternatives. DOLCE was subsequently chosen as a fairly compact and flexible ontology. DOLCE is criticized by Smith & Ceusters (2010) because it expands beyond strict ontological realism into "possible worlds", such as mythology or fiction. This however is seen as a justification for the use of

[2] GitHub - ucoProject/UCO: This repository is for development of the Unified Cyber Ontology.

[3] Ontospy > uco-1.2.0-docs (unifiedcyberontology.org)

[4] Ontospy > uco-1.2.0-docs (unifiedcyberontology.org)

it when describing a manmade and partially metaphorical domain. The conclusion is supported by Semy, et al. (2004).

The used CRA methodology bears close resemblance to the Design Science Research (DSR) methodology, which is widely used in Information Systems Research (Piirainen & Gonzalez, 2014). The authors also present an analysis of the differences in their article (Piirainen & Gonzalez, 2014). According to the analysis, DSR and CRA have many similarities, but DSR puts more weight on supporting the design on an earlier kernel theory than CRA, which is explicitly based on a pragmatist research philosophy. Continuing, the evaluation is more extensive in DSR. The methodology of CRA as described by Lehtiranta, et al. (2015) is based on six phases. In these, (1) the problem is defined, (2) an understanding of the topic is generated, (3) a solution (model) is constructed which then is (4) demonstrated. Then the models (5) theoretical connections are presented and the (6) scope of applicability is assessed.

3. The Constructive Research Approach

The Constructive Research Approach (CRA), presented by Kasanen, et al. (1993) was refined for project management by Lehtiranta, et al. (2015). The first and the second phase of the methodology were in part concluded in an earlier paper (Grandin, 2023). The paper was also the first step in obtaining an understanding of the subject through a systematic literature review. The understanding is here further developed towards ontologies of cyberspace and cybersecurity, as well as ontology engineering. The WebProtégé tool was chosen to construct the ontology. It is a lightweight ontology editor that supports collaboration and has a user-friendly interface (Tudorache, et al., 2013).

The third step of the CRA is designing the constructs. The extensive literature review of cyberspace related ontologies was carried out using Google Scholar. In a recent systematic literature review, almost all references in this area of research were found using it (Martins, et al., 2020). The extensive literature review used the search terms “Cyberspace Ontology” and “Cybersecurity Ontology” and was extended based on the results. Results after the year 2000 were included and resulted in 53 articles that were chosen after analysis of the abstract. In the literature review no core or reference ontology of cyberspace, as defined here, was found. Instead, multiple domain ontologies incorporating either different aspects of cybersecurity or a combination of cybersecurity and decision-making or situational awareness were discovered.

Since cyberspace is metaphorical and contextual, the Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) foundational ontology was selected as a starting point for the work. The DOLCE ontology is descriptive and has a clear cognitive bias (Masolo, et al., 2003). DOLCE is available in multiple versions. According to the homepage of the laboratory maintaining DOLCE, the OWL versions “*do not cover modality, nor temporal indexing, and include the representation of entities such as descriptions and situations which are not covered in DOLCE* “[5]. The Dolce+DnS Ultralite version (DUL)[6] was chosen for the work, as it is a simplification that also include a descriptions and situation ontology (DnS). A simplified picture of the main classes of DUL is shown in figure 2 below.

[5] Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE) – Laboratory for Applied Ontology (LOA) (cnr.it)

[6] Ontology:DOLCE+DnS Ultralite - Odp (ontologydesignpatterns.org)

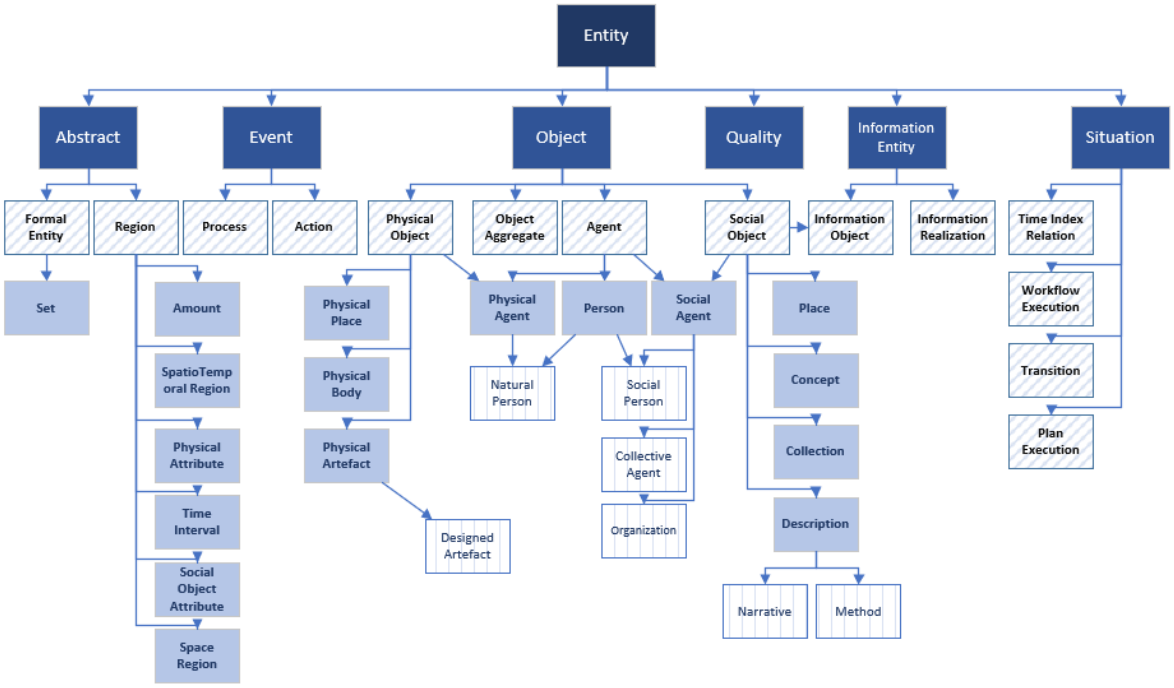


Figure 2: A simplification of the uppermost DUL classes are shown in shades of blue

The design of CoCy begun by linking the earlier mentioned six levels of cyberspace with DUL. Then the core ontology would be linked to a cybersecurity domain ontology, as proof of concept. For this, the earlier mentioned UCO ontology was chosen as a comprehensive attempt to link cybersecurity application ontologies and standards. UCOThing is the core class of the top classes presented in figure 1. The main subclasses of UCOThing are UCOInherentCharacterizationThing and UCOObject. According to the ontology documentation, the first mentioned is "*a grouping of characteristics unique to a particular inherent aspect of a UCO domain object.*". Continuing, UCOObject contains fundamental concepts of the cyber domain and functions as a base class for all content objects. The top classes of UCOThing is shown in figure 3 below.

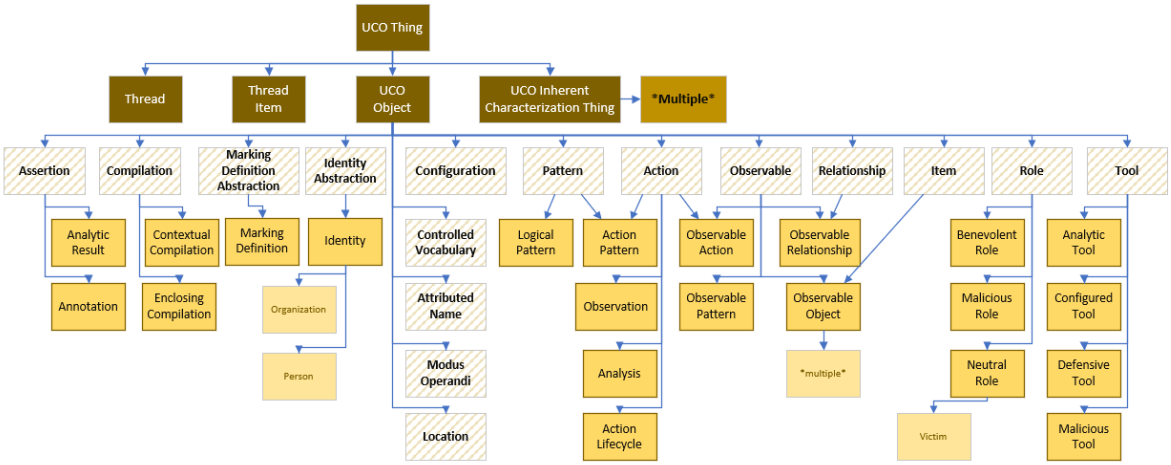


Figure 3: The subclasses of the UCOThing class, focusing on UCOObject

Since the work was identified as laborious, some restrictions were defined. Firstly, only the classes of the ontologies were to be linked and properties and relations only later. Secondly, only the UCOObject class and its subclasses was chosen as sufficient proof of concept. The work was carried out by creating the classes and the links between them in Office Visio.

4. An Ontology of Cyberspace

The six-level starting point of this paper linked DUL and partially the UCO ontologies. Cyberspace is herein seen as a digital extension of the world, and linked to DUL through the Concept class. The levels persona and

geographical information were left unlinked to Cyberspace, since the concepts were sufficiently described in DUL and would not directly benefit from a link to the Cyberspace class. CoCy was also extended in the process. Mainly, there was a need to describe information entities and action in cyberspace. This resulted in three new classes under Cyberspace, linked directly to corresponding classes in DUL. These are Action, InformationRealization and InformationObject. The following two classes from UCO were directly linked to DUL. ModusOperandi was linked to the class Method and Role to the class Narrative. These could be defined as double-linked to the class Cyberspace, but the subject needs further examination. The subclasses of the levels are still under construction but includes at the moment the most basic entities of the classes. The resulting core ontology is shown in figure 4.

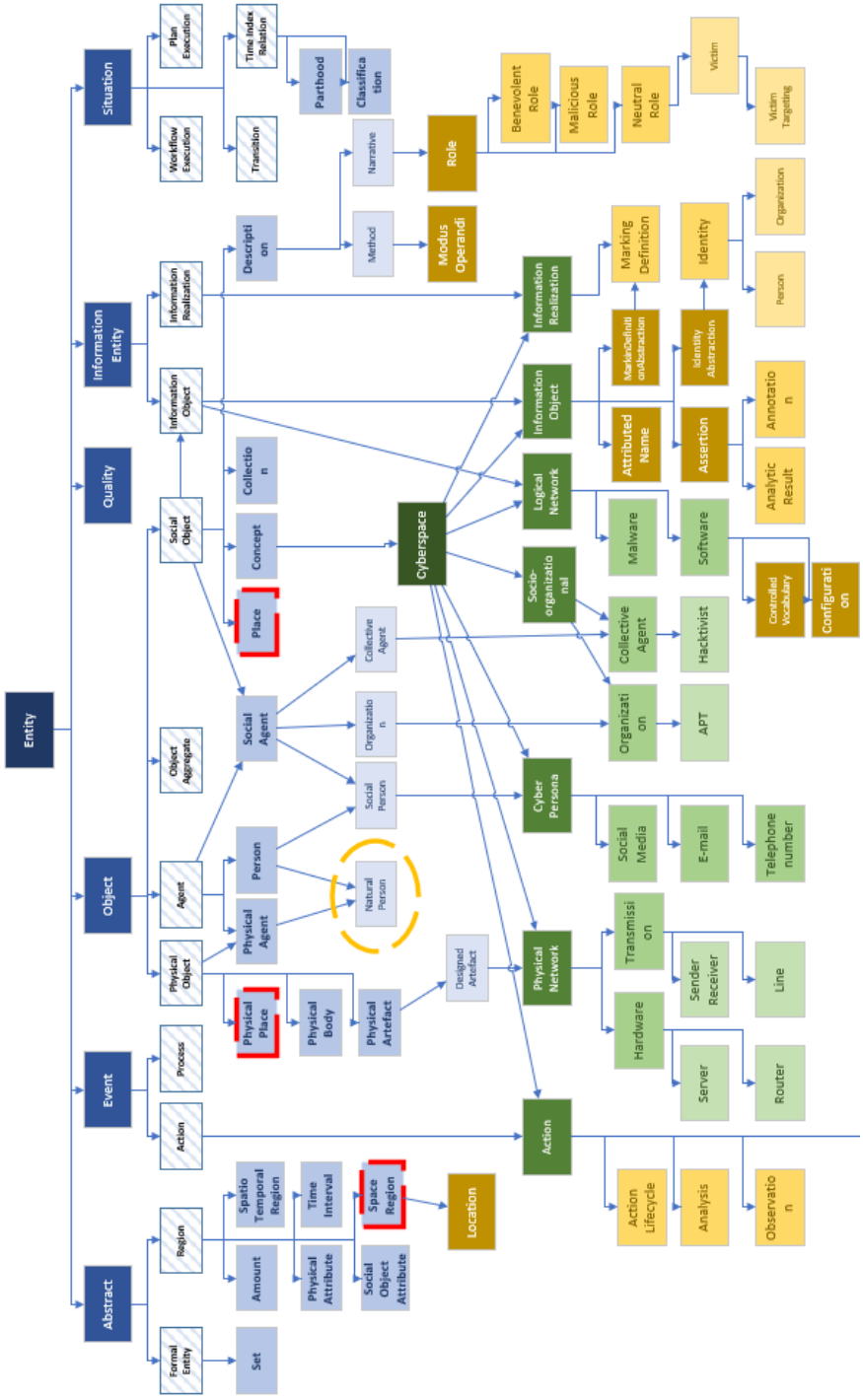


Figure 4: The ontology classes of DUL shown in blue, cyberspace in green and UCO in brown. The classes highlighted with a dashed line are the planes not directly linked to CoCy but essential to cyberspace planning

5. Challenges in Creating a Core Ontology Linking UCO to DUL

5.1 Challenges Linking CoCy to DUL

The connection between DUL and the defined levels of cyberspace encountered some challenges. Primarily the difficulties reside in the concept of cyberspace as a digital imprint of the physical world. A reference ontology (here also seen as a requirement for core ontologies) should according to Rudnicki, et al. (2016) provide all necessary nodes to support the development of lower-level ontologies. Cyberspace consist of parts that could form domain ontologies of their own, e.g. the physical network, cyber persona and the logical network. The guidelines they present are thus challenging to adhere to. E.g. in CoCy the PhysicalNetwork ends in a root node that is linked to the DUL class DesignedArtifact, CyberPersona to the DUL class SocialPerson and the LogicalNetwork to the DUL class InformationObject. To keep these linked to the core ontology, the class Cyberspace was constructed, which is a subclass of the DUL class Concept. The mentioned levels thus have two links – to the common world through the DUL class and through cyberspace relevance to the Cyberspace class. This is depicted in figure 5 below.

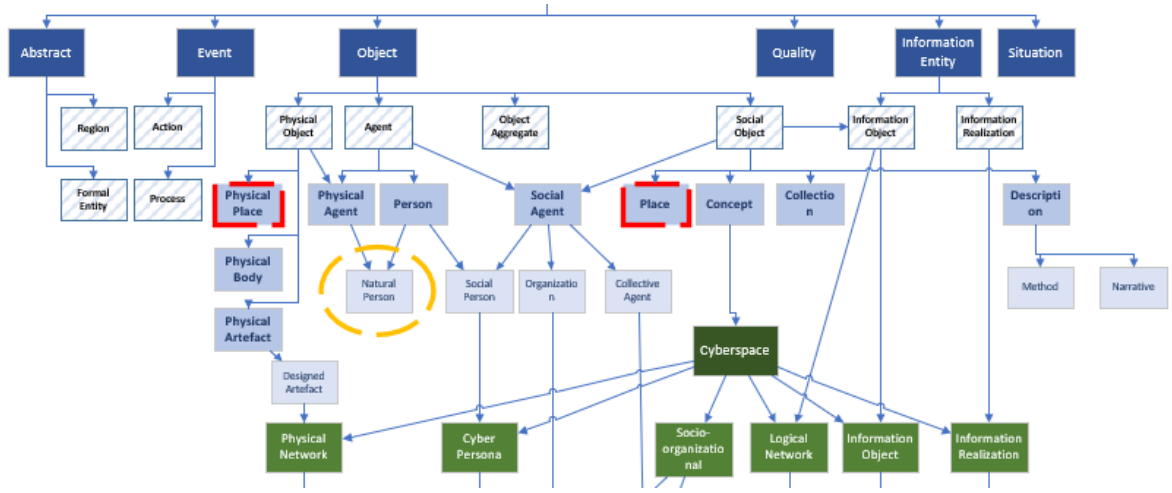


Figure 5: DUL classes are shown in shades of blue and CoCy classes in green. Geographical information and natural persons are directly linked to DUL classes marked with a dashed line. In the middle the DUL Concept subclass Cyberspace which acts as a unifying class

The created Cyberspace class denotes entities that are relevant to cyberspace. The union of the constructed subclasses and the class Cyberspace thus discards the entities that are irrelevant for the core ontology, but could otherwise be included in the description of the class (e.g. cyberspace irrelevant organizations).

Most DUL classes have a digital coupling which make them relevant for CoCy. For instance, the geographical plane (1) and the persona plane (6), are essential for operational planning regarding cyberspace. It is worth noting that these, as defined here, lack a link to the Cyberspace class. These levels present information that exist independently of cyberspace and are thus directly defined by DUL classes. Human persons are thus linked to the class NaturalPerson in DUL and geographical information to classes AbstractRegion, PhysicalPlace and Place in DUL. These can also be regarded as extensions of the "core" of cyberspace, defined in planes (2) to (5). In for instance a wartime situation a military commander needs the information how an attack through cyberspace is linked to physical assets and persons to determine the best course of action. As such, the information presented to a decision maker should include the options to engage either the geographical or persona planes if assessed efficient. The importance of this information is for instance exemplified in Michel & King (2019). Another use case is identifying and apprehending cybercriminals or to prevent cybercrime. The link between the geographical plane and DUL and also physical persons is depicted in figure 6 below. As seen, the core ontology would be difficult to link to a single root node and still be grounded in the DUL foundational ontology. Hence, CoCy is a conceptual extension of DUL and will contain all of DUL and also the extensions of cyberspace classes.

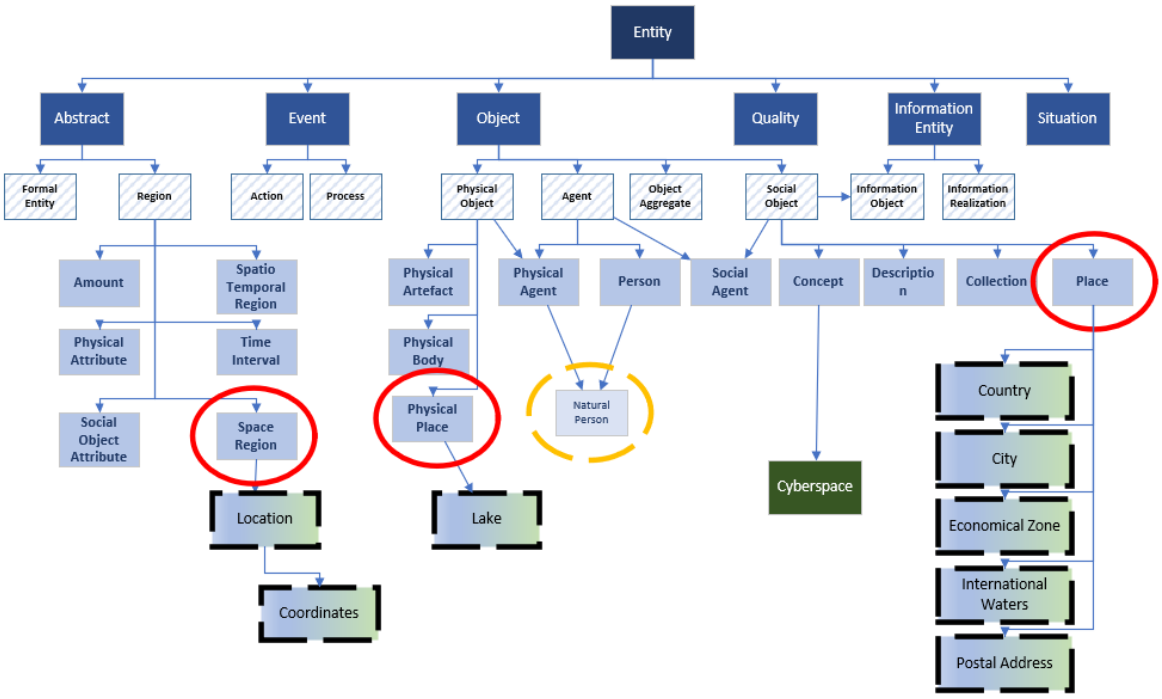


Figure 6: A simplified picture of the classes that are relevant, but not linked to the created Cyberspace class. Geographical information exists in the DUL classes highlighted with a red ellipse. Preliminary sub-classes are highlighted with a dashed black line. Natural persons are grouped under the DUL class Natural person, highlighted with a yellow dashed ellipse

5.2 Challenges Linking CoCY to UCO

Early on, it became clear how the construction of CoCy and UCO differed. DUL and CoCy are created top-down as where the UCO is a compilation of several existing taxonomies, information exchange and categorization standards, and constructed bottom-up. UCO also showed to be constructed mainly according to the concept of combining similar entities into classes with a (cybersecurity) technical point of view and not according to the ontological realism of a foundational ontology. For instance, the UCOObject subclass ObservableObject includes a large variety of cybersecurity related characteristics, such as software, information, devices etc. as shown in figure 7 below. The class is in the documentation described as "*a grouping of characteristics unique to a distinct article or unit within the digital domain*.".

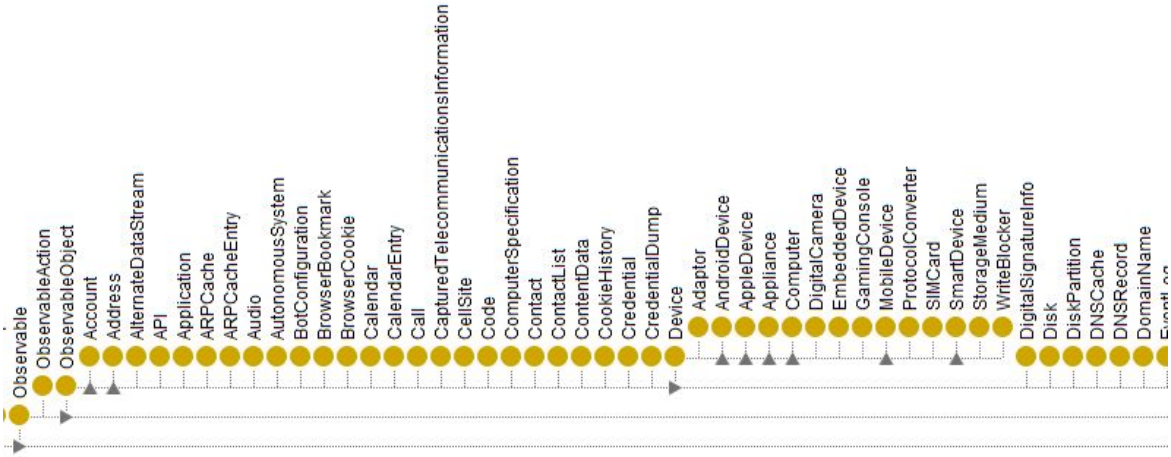


Figure 7: A part of the subclasses of the UCO ObservableObject class, showing information, devices and accounts etc.

These would in DUL be subclasses of InformationRealization, InformationObject, DesignedArtefact or SocialPerson (accounts) etc. While many of the classes of UCO are possible to link either directly to DUL or CoCy classes, the above-mentioned class presents a challenge. In essence, it would be practical if the

ObservableObject class of UCO would be divided into multiple classes that differentiated between the different physical aspects of the sub-classes. Another problematic example are classes including hardware and software components, such as the subclass Tool, which include instances of as well InformationRealization, InformationObject and DesignedArtefact. The upper level classes from UCO that were unsuccessfully linked are shown in figure 8 below.

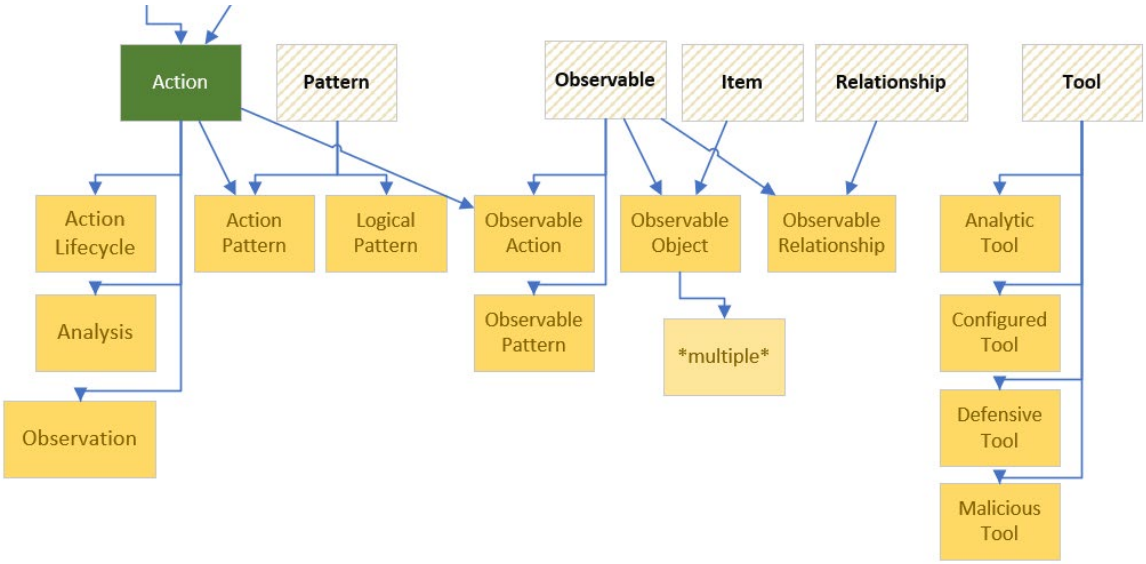


Figure 8: The classes of UCO that were unsuccessfully linked to either DUL or CoCy are shown in brown. The CoCy class Action is included to depict interconnections

The linking of CoCy to UCO also revealed important aspects of cyberspace undefined by the original six planes of cyberspace, namely information (entities) and actions. These were included in CoCy. Cyberspace-related information is linked to the class Cyberspace and either to the general DUL InformationObject and InformationRealization classes.

6. Conclusions

This paper presents the ongoing work of constructing a core ontology for cyberspace called CoCy. The domain lacks a unifying core ontology, and this paper has demonstrated that such an ontology can be constructed, grounded to a foundational ontology. Since cyberspace is a domain that imbues the world we live in, and as such consists of multiple sub-domains, the core ontology was created as an extension of the DUL foundational ontology. The starting point of the work, which was a model of cyberspace consisting of six different planes, evolved by excluding the periphery levels persona and geographical information. The exclusion was done by linking this information directly to DUL. The model was extended by including classes of action and information. Due to the fundamental differences of the foundational ontology and the chosen proof of concept (the Unified Cyber Ontology UCO), some classes were unsuccessfully mapped to the core ontology. These difficulties are however manageable, but requires further research.

The proposed solution in form of an extension of DUL needs further research on its feasibility. Also, the introduction of the class Cyberspace should be further investigated. The work on writing the proposed core ontology in OWL using WebProtégé, continues. As such, research adhering to the chosen methodology still needs to 1) be demonstrated, 2) present theoretical connections and finally, 3) continue by assessing the scope of applicability. The work with linking CoCy to UCO however validated the need for a flexible core ontology for cyberspace, grounded in a foundational ontology.

References

Asquith, P. & Morgan, P., 2020. Representing a Human-Centric Cyberspace. In: I. Corradini, E. Nardelli & T. Ahram, eds. *Advances in Human Factors in Cybersecurity, Advances in Intelligent Systems and Computing, vol 1219.* Cham: Springer, pp. 122-128.

Babayeva, G., Maennel, K. & Maennel, O. M., 2022. *Building an Ontology for Cyber Defence Exercises.* Genoa, IEEE, pp. 423-432.

Bowman, M., Lopez, A. & Tecuci, G., 2001. *Ontology Development for Military Applications.* New York, ACM Press.

Bromander, S. et al., 2020. Modeling Cyber Threat Intelligence. *Proceedings of the 6th International Conference on Information Systems Security and Privacy (ICISSP 2020),* pp. 273-280.

Burger, E. W., Goodman, M. D., Kampanakis, P. & Zhu, K. A., 2014. *Taxonomy Model for Cyber Threat Intelligence Information Exchange Technologies.* New York, Association for Computing Machinery, pp. 51-60.

Chan, P., Theron, J., van Heerden, R. & Leenen, L., 2015. *An ontological knowledge base for cyber network attack planning.* Kruger National Park, South Africa, Academic Conferences Limited, pp. 69-77.

Dipert, R. R., 2013. The essential features of an ontology for cyberwarfare. In: *Conflict and Cooperation in Cyberspace : The Challenge to National Security.* Boca Raton: CRC Press, pp. 35-48.

Grandin, A., 2023. *Cyber Geography and Cyber Terrain: Challenges producing a universal map of Cyberspace.* Reading, UK, Academic Conferences International Limited, pp. 207-213.

Grant, T. J., 2014. Formalized Ontology for Representing C2 Systems as Layered Networks. In: T. J. Grant, R. H. P. Janssen & H. Monsuur, eds. *Network Topology in Command and Control: Organization, Operation, and Evolution.* s.l.:IGI Global, pp. 85-124.

Huntley, W. L., 2016. *Cyber Key Terrain: A Conceptual Assessment,* s.l.: U.S. Naval Postgraduate School.

Iannacone, M. et al., 2015. *Developing an Ontology for Cyber Security Knowledge Graphs.* New York, Association for Computing Machinery, pp. 1-4.

Joint Chiefs of Staff, 2018. *Joint publication 3-12, Cyberspace Operations,* Washington DC: Joint Chiefs of Staff.

Kasanen, E., Lukka, K. & Siitonen, A., 1993. The constructive approach in management accounting research. *Journal of Management Accounting Research,* pp. 243-264.

Kreuzer, M. P., 2021. *Cyberspace is an analogy, not a domain: Rethinking domains and layers of warfare for the information age.,* s.l.: The Strategy Bridge.

Lehtiranta, L., Junnonen, J.-M., Kärnä, S. & Pekuri, L., 2015. The Constructive Approach: Problem Solving for Complex Projects. In: *Designs, Methods and Practices for Research of Project Management.* s.l.:Gower Applied Business Research, pp. 95-106.

Maathuis, C., Pieters, W. & van den Berg, J., 2018. Developing a Cyber Operations Computational Ontology. *Journal of Information Warfare , Vol. 17, No. 3,* pp. 32-49.

Martins, B. F. et al., 2022. A framework for conceptual characterization of ontologies and its application in the cybersecurity domain. *Software and Systems Modeling volume 21,* pp. 1437-1464.

Martins, B. et al., 2020. *Conceptual Characterization of Cybersecurity Ontologies.* s.l., Springer, pp. 323-338.

Mascardi, V., Cordì, V. & Rosso, P., 2007. *A Comparison of Upper Ontologies,* s.l.: Workshop From Objects to Agents.

Masolo, C. et al., 2003. *Wonderweb deliverable d18: Ontology library,* s.l.: Laboratory for Applied Ontology, ISTC-CNR.

Matheus, C. J. & Ulicny, B., 2007. *On the Automated Generation of an OWL Ontology based on the Joint C3 Information Exchange Data Model.* s.l., s.n., pp. 1-19.

Mavroeidis, V. & Bromander, S., 2017. *Cyber Threat Intelligence Model: An Evaluation of Taxonomies, Sharing Standards, and Ontologies.* s.l., IEEE, pp. 91-98.

Michel, M. C. K. & King, M. C., 2019. *The Future of Cyber Analytics: Identity Classification for Systematic and Predictive Insight.* Oxford, UK, IEEE, pp. 1-4.

Morosoff, P. et al., 2015. *Joint Doctrine Ontology: A Benchmark for Military Information Systems Interoperability..* s.l., CEUR-WS, pp. 2-9.

Möller, D. P. F., 2020. Cybersecurity Ontology. In: *Cybersecurity in Digital Transformation : Scope and Applications.* Cham, Switzerland: Springer, pp. 99-109.

NATO, 2020. *Allied Joint Publication-3.20 Allied Joint Doctrine for Cyberspace Operations,* s.l.: NATO STANDARDIZATION OFFICE (NSO) .

Obrst, L., Chase, P. & Markeloff, R., 2012. *Developing an Ontology of the Cyber Security Domain..* s.l., CEUR-WS, pp. 49-56.

Oltramari, A., Cranor, L. F., Walls, R. J. & McDaniel, P. D., 2014. Building an Ontology of Cyber Security. *Semantic Technologies for Intelligence, Defense, and Security,* pp. 54-61.

Pai, F.-P., Yang, L.-J. & Chung, Y.-C., 2017. Multi-layer ontology based information fusion for situation awareness. *Applied Intelligence 46,* pp. 285-307.

Phister, P. W., 2010. Cyberspace: The Ultimate Complex Adaptive System. *The International C2 Journal | Vol 4, No 2,* pp. 1-25.

Piirainen, K. A. & Gonzalez, R. A., 2014. Constructive Synergy in Design Science Research: A Comparative Analysis of Design Science Research and the Constructive Research Approach. In: *Liiketaloudellinen aikakauskirja LTA - Finnish Journal of Business Economics.* s.l.:Liiketaloustieteellinen Yhdistys ry, pp. 206-234.

Rudnicki, R., Smith, B., Malyuta, T. & Mandrick, W., 2016. *Best Practices of Ontology Development,* s.l.: NIST.

Semy, S. K., Pulvermacher, M. K. & Obrst, L. J., 2004. *Toward the Use of an Upper Ontology for U.S. Government and U.S. Military Domains: An Evaluation,* Bedfod, Massachusetts: MITRE.

Sikos, L. F., 2018. OWL Ontologies in Cybersecurity: Conceptual Modelling of Cyber-Knowledge. In: *AI in Cybersecurity.* Cham, Switzerland: Springer, pp. 1-17.

Sikos, L. F., 2023. Cybersecurity knowledge graphs. *Knowledge and Information Systems 65,* p. 3511–3531.

Smith, B. & Ceusters, W., 2010. Ontological realism: A methodology for coordinated evolution of scientific ontologies. *Applied Ontology, vol. 5, no. 3-4,* pp. 139-188.

Syed, Z. et al., 2016. *UCO: A Unified Cybersecurity Ontology.* Baltimore, University of Maryland, Baltimore County.

Takahashi, T. & Kadobayasi, Y., 2015. Reference ontology for cybersecurity operational information. *The Computer Journal vol. 58 no.10,* pp. 2297-2312.

Tudorache, T., Nyulas, C., Noy, N. F. & Musen, M. A., 2013. WebProtégé: A collaborative ontology editor and knowledge acquisition tool for the web. *Semantic web, 4(1),* pp. 89-99.

van Heerden, R. C. P., Leenen, L. & Theron, J., 2016. *Using an Ontology for Network Attack Planning.* s.l., IGI Global, pp. 65-78.

Learn, Unlearn and Relearn: Adaptive Cybersecurity Culture Model

Tapiwa Gundu
Nelson Mandela University, Centre of Research in Information and Cyber Security (CRICS), Gqeberha, South Africa

tapiwag@mandela.ac.za

Abstract: In the ever-evolving cyberspace landscape, organisations face persistent threats that continuously mutate and adapt. To effectively defend against these dynamic cyber threats, a fundamental shift in cybersecurity culture is imperative. This paper presents a novel Adaptive Cybersecurity Culture Model (ACCM) that encapsulates the principles of "Learn, Unlearn, and Relearn" as a strategic stance to foster resilience and adaptability in the face of evolving cyber threats. The ACCM emphasizes the importance of continuous learning as the cornerstone of cybersecurity culture. It advocates the adoption of a growth mindset within organisations, encouraging employees to stay updated with emerging threats, technologies, and best practices. However, the model goes beyond mere learning; it underscores the significance of unlearning outdated practices and misconceptions that may hinder effective cybersecurity. Furthermore, the ACC model introduces the concept of "Relearn," emphasizing the need to rapidly adapt and evolve strategies and tactics in response to ever-changing cyber threats. It promotes a culture of agility and adaptability, enabling organisations to respond effectively to both known and unforeseen cyber challenges. The model presented in this paper draws from case studies of various industries to illustrate the successful adoption and transformation of cybersecurity culture using the Learn, Unlearn, and Relearn principles. The ACCM represents a paradigm shift in cybersecurity culture, acknowledging that the ability to adapt is as critical as the ability to protect. By fostering a culture of continuous learning, unlearning, and relearning, organisations can proactively enhance their cyber resilience and effectively defend against the ever-evolving cyber threat landscape. This paper provides a roadmap for organisations to embark on this transformative journey toward a more adaptive and resilient cybersecurity culture.

Keywords: Cybersecurity culture, Adaptive cybersecurity, Learn, Unlearn, Relearn, Cyber threat resilience, Growth mindset, Continuous learning

1. Introduction

Cybersecurity has become a paramount concern in the digital age due to the increasing dependence on digital technologies and the proliferation of interconnected systems (Gundu, 2023a). The rapid growth of the internet, cloud computing, mobile devices, and Internet of Things (IoT) has brought about unprecedented opportunities for innovation, communication, and convenience (Gundu et al., 2019). However, it has also given rise to a host of complex and evolving cyber threats that pose significant risks to individuals, organisations, and even nations (Craig, 2018). The digital age is characterized by a widespread transformation of various sectors, including finance, healthcare, energy, transportation, and communication. This transformation is driven by the adoption of digital technologies, data-driven decision-making, and the interconnectivity of devices and systems (Jin et al., 2023). While this connectivity enhances efficiency and accessibility, it also exposes critical assets and sensitive information to potential cyberattacks (Gundu, 2023b).

Massive data breaches have become common place, compromising the personal and financial information of millions of individuals (Madan et al., 2023). High-profile incidents involving renowned companies (e.g. T Mobile, Chat GPT, MailChimp, Google Fi, etc) have underscored the urgent need for robust cybersecurity measures (Farrelly, 2023). Additionally, concerns about data privacy, surveillance, and the misuse of personal information have led to the development of stringent data protection regulations, such as the European Union's General Data Protection Regulation (GDPR) (Team, 2020), the California Consumer Privacy Act (CCPA) (Wong et al., 2023) or South African Protection of Personal Information Act (POPIA) (Netshakhuma, 2019). The consequences of cyberattacks extend beyond immediate financial losses. Cyber incidents can disrupt critical infrastructure, disrupt supply chains, and result in the loss of intellectual property (Mmango & Gundu, 2023).

As technology advances, so do the tactics, techniques, and procedures of cybercriminals, hacktivists, state-sponsored actors, and other malicious entities. The threat landscape has evolved beyond traditional viruses and malware to include sophisticated attacks such as ransomware, advanced persistent threats (APTs), social engineering, and zero-day exploits (Anisetti et al., 2020; Rahman et al., 2023; Schiller et al., 2022). The ever-changing threat landscape and evolving nature of cybersecurity challenges faced by organisations in the digital realm are evident and can be seen by the continuous mutation and adaption of the threats presenting a moving target for cybersecurity defences. This dynamic environment is characterized by the rapid development of new attack techniques, exploitation methods, and vulnerabilities. Cyber adversaries constantly innovate and modify their tactics to bypass existing security measures, making it essential for organisations to stay vigilant and

responsive. The ever-changing threat landscape underscores the need for a proactive and adaptive cybersecurity culture that goes beyond static defences, emphasizing continuous monitoring, learning, and adjustment to effectively counter emerging cyber threats.

Cybersecurity culture refers to the shared values, beliefs, attitudes, behaviours, and practices within an organisation that collectively shape its approach to information security. Cybersecurity culture plays a pivotal role in determining how an organisation perceives, prioritises, and addresses security issues. A strong cybersecurity culture should foster a proactive and security-conscious mindset among employees at all levels, encouraging them to consider security in their day-to-day activities. It involves creating an environment where individuals understand the importance of protecting sensitive information, recognise potential security risks, and actively contribute to maintaining a secure digital ecosystem. A positive cybersecurity culture promotes awareness, accountability, and a commitment to best practices, ensuring that security measures are not just seen as technical requirements but as integral components of the overall organisational ethos. Cultivating a cybersecurity culture involves leadership commitment, continuous education, clear communication, and the integration of security considerations into the organisation's policies and processes. Given the high stakes associated with cyber threats, a robust cybersecurity culture is instrumental in fortifying an organisation's defences against the ever-evolving landscape of cyber threats.

The objectives of this study are to address the critical need for a paradigm shift in cybersecurity culture within organisations, given the persistent and ever-evolving threats in cyberspace. The primary goal is to introduce and elucidate a novel model, the Adaptive Cybersecurity Culture Model (ACCM), built on the principles of "Learn, Unlearn, and Relearn." This model aims to serve as a strategic baseline for fostering resilience and adaptability in the face of dynamic cyber threats. The study seeks to emphasize the significance of continuous learning as the foundation of a robust cybersecurity culture, advocating for a growth mindset among employees to stay abreast of emerging threats, technologies, and best practices. Furthermore, the study aims to highlight the importance of unlearning outdated practices and misconceptions that may hinder effective cybersecurity measures. The concept of "Relearn" is introduced to underscore the need for organisations to rapidly adapt and evolve strategies in response to ever-changing cyber threats.

2. Related Literature

2.1 Evolution of Cyber Threats and Their Impact on Traditional Security Practices

The evolution of threat vectors in the realm of cybersecurity reflects the dynamic nature of the digital landscape and the continuous innovation by malicious actors seeking to exploit vulnerabilities (Anisetti et al., 2020). Initially, cybersecurity threats were primarily virus and malware-driven, often spread through infected files or email attachments. As technology advanced, so did the threat vectors, with the emergence of more sophisticated tactics such as phishing, where attackers manipulate individuals into divulging sensitive information (Rahman et al., 2023). The rise of interconnected systems and the Internet of Things (IoT) introduced new threat vectors, allowing attackers to exploit vulnerabilities in interconnected devices and networks (Gundu & Maronga, 2019). Social engineering became increasingly prevalent, leveraging psychological manipulation to deceive individuals or gain unauthorized access (Kilavo et al., 2023). Advanced Persistent Threats (APTs) marked another shift, involving prolonged and targeted attacks for data exfiltration or system compromise (Quintero-Bonilla & Martín del Rey, 2020). With the expansion of cloud computing and mobile technologies, threats evolved to exploit weaknesses in these platforms. Additionally, the use of artificial intelligence and machine learning by both defenders and attackers added a layer of complexity to the threat landscape (Taddeo et al., 2019). As technologies like 5G, edge computing, and quantum computing emerge, the evolution of threat vectors continues, requiring cybersecurity strategies to adapt to these advancements and anticipate future challenges. The ongoing evolution underscores the importance of proactive measures, a holistic cybersecurity approach and continuous vigilance by employees (Sun et al., 2023).

2.2 Employees - The Weakest Link

Employees are often considered the weakest link in cybersecurity, highlighting the vulnerabilities introduced by their behaviours, actions, and decisions within the organisations digital environment (Thomson et al., 2006). Despite advancements in technology and sophisticated cybersecurity measures, employees remain susceptible to various cyber threats (Zimmermann & Renaud, 2019). Employee-related vulnerabilities manifest in actions such as falling victim to phishing attacks, using predictable passphrases, neglecting software updates, or unintentionally disclosing sensitive information. One key reason for this vulnerability is the lack of awareness about cybersecurity risks and best practices (Zwilling et al., 2022). Many individuals may not fully comprehend

the potential consequences of their actions in the digital realm, leading to inadvertent security lapses (Moustafa et al., 2021). Cybercriminals often exploit this lack of awareness through social engineering tactics, such as tricking individuals into clicking malicious links or sharing confidential information (Thomson & Niekerk, 2018).

Building cybersecurity awareness is crucial to mitigating the human factor as a weak link (Zwilling et al., 2022). Education and training programs can empower individuals to recognize and respond to cyber threats effectively. This includes understanding the tactics employed by cybercriminals, recognizing phishing attempts, and adopting secure online practices. Awareness efforts should also emphasize the importance of maintaining strong passphrases, use of onetime pins, regularly updating software, and exercising caution in online interactions. However, most information security awareness programs only place strong emphasis on learning new practices, strategies, and skills to enhance an individual's ability to safeguard digital assets and sensitive data neglecting a crucial aspect of unlearning outdated or ineffective security practices (Hielscher et al., 2022).

2.3 The Concept of Unlearning and Relearning in Information Security Awareness and Training Contexts

In the context of information security awareness and training, the concepts of unlearning and relearning play crucial roles in adapting to the ever-evolving cybersecurity landscape. Unlearning involves letting go of outdated practices, misconceptions, or habits that may pose security risks (Zhao & Wang, 2020). In the rapidly changing field of information security, clinging to outdated knowledge or approaches can be detrimental. Therefore, individuals need to actively unlearn practices that may have been effective in the past but are no longer sufficient in addressing current cybersecurity threats.

Relearning, on the other hand, emphasizes the continuous acquisition of new knowledge and skills to stay abreast of emerging threats and technologies (Zhao & Wang, 2020). Cybersecurity is a dynamic field, and what may be considered best practices today might become obsolete tomorrow. Relearning involves staying proactive, seeking out the latest information on cybersecurity threats, tools, and strategies, and adapting one's knowledge base to effectively counter new and sophisticated challenges.

Awareness campaigns should not only focus on imparting foundational knowledge but also encourage participants to question existing practices, discard obsolete methodologies, and adopt new and more effective security measures (Sharma & Lenka, 2019). This involves creating a culture of agility and adaptability, where individuals are empowered to evolve their understanding of cybersecurity in response to the rapidly changing threat landscape. Moreover, case studies and real-world examples of security breaches can serve as powerful tools for unlearning and relearning. Analysing past incidents allows individuals to understand the consequences of outdated security practices and reinforces the importance of staying current with the latest cybersecurity trends.

2.4 Adaptive Cybersecurity

Adaptive cybersecurity involves the continuous process of adjusting and evolving cybersecurity strategies, practices, and technologies to effectively counter new and emerging threats (Addae et al., 2017). Unlearning and relearning play a pivotal role in achieving adaptive cybersecurity by addressing outdated practices and acquiring new knowledge and skills. For ease of adaptive learning, people should possess cognitive skills, these involve the mental capacity to switch between different concepts, perspectives, or strategies in response to changing situations (Caulkins et al., 2019).

2.5 Cognitive Flexibility and its Relevance to Adaptive Cybersecurity

Cognitive flexibility refers to the ability of employees to mentally adapt, shift perspectives, and adjust their thinking to effectively respond to evolving cyber threats. It is a cognitive skill that enables employees to navigate the complex and dynamic nature of the cybersecurity landscape .Cognitive flexibility enables the employees to adapt to new information, think creatively, and solve complex problems (Caulkins et al., 2019). They can shift their focus and strategies to address emerging threats without being constrained by rigid mindsets or established practices.

In an ever-changing cyber landscape, cognitive flexibility enables professionals to identify anomalies and patterns that may indicate a potential attack (Ask et al., 2023). Cyber threats are often not straightforward and may involve complex, non-obvious indicators. During a cyber incident, the ability to switch between different response strategies and adapt to the changing nature of the attack is crucial. Cognitive flexibility helps cybersecurity teams quickly assess the situation, determine the appropriate course of action, and adjust their response as needed.

Ultimately, cognitive flexibility supports a culture of continuous learning and improvement within cybersecurity teams (Caulkins et al., 2019). Professionals who can readily unlearn outdated practices and relearn new ones contribute to the organisation's ability to stay ahead of emerging threats.

2.6 Factors Contributing to the Persistence of Outdated Practices

The persistence of outdated security practices within organisations can be attributed to a combination of factors. Key contributing factors include cognitive inertia, as people tend to resist change and may be reluctant to unlearn outdated practices; lack of awareness about the rapid evolution of cyber threats; resource constraints, with limited budgets hindering investments in new technologies; organisational culture favouring tradition over innovation; misaligned incentives when individuals are unaffected by outdated practices; fear of disruption during the transition to new practices; compliance-driven decision-making that can impede adaptive change; a lack of training and skill development; the perception of low risk in the absence of significant incidents; unclear communication about the urgency of updating security practices; a short-term focus prioritizing immediate tasks over long-term cybersecurity preparedness; and the challenge of legacy systems and dependencies. Overcoming these factors necessitates a comprehensive approach involving effective communication, education, leadership support, incentivisation, and a cultural shift towards adaptability and innovation, enabling organizations to successfully unlearn outdated practices and embrace the principles of adaptive cybersecurity.

2.7 Risks Associated With Clinging to Obsolete Security Practices

Clinging to outdated cybersecurity practices poses numerous risks and adverse consequences for organisations. Such practices increase vulnerability to emerging threats, leaving organizations exposed to modern attack techniques and advanced persistent threats (Sharma & Lenka, 2019). Ineffective threat detection and response capabilities may result in delayed incident responses and prolonged unauthorized access. This can lead to data breaches, compromising sensitive information and exposing organizations to legal consequences and non-compliance with data protection regulations. Cybersecurity incidents stemming from obsolete practices can disrupt business operations, impact revenue, and harm customer trust and loyalty. Moreover, organizations adhering to outdated practices may hinder innovation and competitiveness, impacting employee morale, productivity, and the ability to attract and retain skilled cybersecurity professionals (Zhao & Wang, 2020). To mitigate these risks, organizations must prioritize the adoption of adaptive cybersecurity practices that align with current threat trends, leverage modern technologies, and promote a culture of continuous learning and improvement.

3. The Proposed Model

The Adaptive Cybersecurity Culture Model is cyclical, emphasizing that cybersecurity environment should not be, instead, it requires a continuous, dynamic, and iterative process. The model acknowledges that learning is not enough; unlearning outdated practices and relearning in response to evolving threats are equally critical components of a robust cybersecurity culture. This iterative process ensures that organizations are not only prepared for current threats but also adaptive to future challenges in the ever-changing cybersecurity landscape.

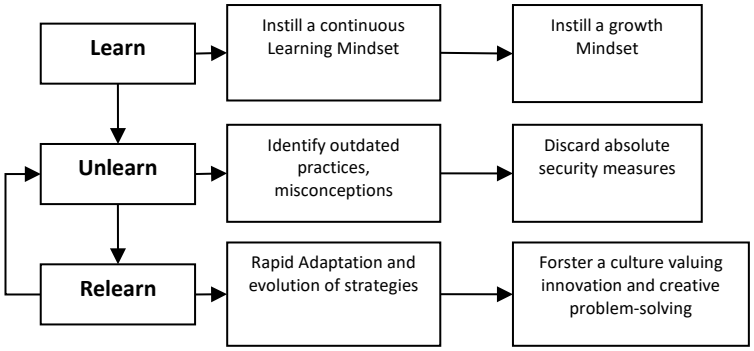


Figure 1: Adaptive Cybersecurity Culture Model

3.1 Learn

Continuous Learning: The "Learn" component emphasizes the importance of ongoing education and development in the field of cybersecurity. This involves implementing regular training programs, workshops, and

educational sessions to ensure that cybersecurity professionals stay informed about the latest threats, technologies, and best practices.

Growth Mindset: Encouraging a growth mindset within the organization is crucial. This mindset fosters the belief that skills and intelligence can be developed through dedication and hard work. In the context of cybersecurity, it motivates individuals to embrace challenges, learn from failures, and seek opportunities for improvement.

3.2 Unlearn

Identify Outdated Practices: The "Unlearn" component involves recognizing and acknowledging practices, methodologies, or beliefs that have become outdated or ineffective in the face of evolving cyber threats. This step requires a critical assessment of existing security measures to identify areas that need to be revamped or abandoned.

Willingness to Let Go: Unlearning is not just about recognizing outdated practices but also about fostering a culture where individuals are willing to let go of established habits or approaches that may no longer contribute to effective cybersecurity. It involves overcoming resistance to change and embracing new and improved strategies.

3.3 Relearn

Rapid Adaptation: In the "Relearn" phase, organizations focus on the ability to rapidly adapt and evolve their cybersecurity strategies in response to changing threats. This requires a proactive approach to adjusting security measures, staying ahead of emerging risks, and implementing timely updates to defences.

Foster a Culture of Innovation: Relearning emphasizes fostering a culture that values innovation and creative problem-solving. It encourages cybersecurity professionals to think outside the box, explore new approaches, and continuously improve strategies based on real-world experiences and feedback.

Continuous Improvement: Implementing feedback loops is crucial for continuous improvement. This involves regularly evaluating the effectiveness of security measures, learning from incidents or breaches, and making iterative adjustments to enhance overall cybersecurity posture.

4. Discussion

Relearning new techniques is a crucial component of the Adaptive Cybersecurity Culture Model, involving the continuous acquisition and mastery of updated knowledge, skills, and practices to effectively counter evolving cyber threats. In the dynamic landscape of cybersecurity, keeping abreast of the latest techniques is imperative for maintaining a robust defence against rapidly changing attack vectors and vulnerabilities. One area of focus within adaptive cybersecurity involves emerging employee-related trends and techniques that address the human element of cybersecurity. These strategies acknowledge that employees are both targets and potential vulnerabilities, aiming to enhance user awareness, behaviour, and overall cyber hygiene. Noteworthy trends include the use of behavioural analytics and user profiling to detect anomalies, phishing simulation and testing to assess employees' recognition of social engineering attempts, and continuous user authentication methods like biometrics and multi-factor authentication. Other approaches encompass insider threat detection, secure remote work practices, threat deception, and employee-centric incident response plans. Additionally, organisations should foster cybersecurity cultures that integrate security into daily operations, utilising tailored training, microlearning, ethical hacking programs, and security champions to enhance overall cybersecurity awareness. Recognising the emotional and psychological factors influencing user behaviour, these trends collectively contribute to a more resilient cybersecurity posture, reducing the likelihood of successful cyberattacks stemming from employee vulnerabilities.

Strategies for acquiring new knowledge and skills in cybersecurity are multifaceted and essential for employees to navigate the rapidly evolving landscape of cyber threats. A foundational approach involves the increasing investment in comprehensive security awareness training programs, designed to educate employees about common cyber threats, phishing tactics, and safe online practices, thereby minimizing the risk of errors. An innovative strategy employs gamification for training, integrating gaming elements into cybersecurity education to make the learning process engaging, foster active participation, and enhance knowledge retention among employees. Additionally, efforts are being made to make cybersecurity education more inclusive, addressing diverse learning styles, backgrounds, and abilities. Embracing a continuous learning mindset is emphasized, recognizing the dynamic nature of the cybersecurity field, encouraging individuals to stay proactive in updating their knowledge and skills to align with the latest industry developments. Versatile approaches include enrolling

in online courses, tutorials, and e-learning platforms offered by websites like Coursera, edX, Udemy, and Khan Academy, covering a wide array of cybersecurity topics. Pursuing recognized professional certifications such as CompTIA Security+, CISSP, CEH, or CCSP remains a cornerstone for skill acquisition. Seeking mentorship from experienced professionals, following cybersecurity blogs and podcasts, and cultivating curiosity to explore new topics further contribute to a comprehensive and adaptive framework for acquiring and honing cybersecurity knowledge and skills.

In the pursuit of an adaptive cybersecurity culture, organizations must instill a mindset of unlearning and relearning. Key strategies for fostering such a culture include securing leadership support and buy-in, where organizational leaders play a pivotal role in endorsing continuous learning and motivating employees to discard outdated practices for embracing new knowledge (Thomson et al., 2006). Clear communication and awareness campaigns are essential to underscore the imperative of unlearning and relearning (Bada et al., 2019), emphasizing the rapidly changing nature of cybersecurity threats and the inherent benefits of staying current. Creating diverse learning opportunities through regular training, workshops, and access to educational resources, along with hands-on practice and experimentation, equips employees with the skills needed to adapt to evolving challenges (Anisetti et al., 2020). Leading by example, wherein organizational leadership openly discusses their commitment to ongoing learning, sets a positive precedent for employees to follow suit. Recognizing and rewarding active participation in learning initiatives (Gundu et al., 2019), as well as incorporating learning into performance evaluations, further reinforces the importance of continuous skill development. Encouraging cross-functional collaboration, providing channels for feedback and reflection, and establishing a safe environment for experimentation are additional strategies. Recognising changing paradigms, promoting self-directed learning, and utilizing case studies and lessons learned contribute to a comprehensive approach. Ensuring inclusivity in learning programs, celebrating learning achievements, and fostering a long-term outlook on unlearning and relearning as continuous processes are integral components of creating a culture where adapting to new challenges is not only encouraged but celebrated. Ultimately, these strategies collectively form the foundation for achieving and maintaining excellence in cybersecurity and other dynamic fields.

5. Conclusion

In conclusion, this paper introduces an Adaptive Cybersecurity Culture Model (ACCM) to effectively combat the ever-evolving landscape of cyber threats. ACCM is grounded in the principles of "Learn, Unlearn, and Relearn," emphasising the continuous pursuit of knowledge, shedding outdated practices, and swiftly adapting to emerging threats. The strategies outlined for cultivating this culture encompass leadership support, clear communication, diverse learning opportunities, and recognition of employees' active participation. By fostering a culture that embraces ongoing learning and adaptation, organizations can bolster their resilience and responsiveness in the face of dynamic cybersecurity challenges. The recognition of human factors as central to cybersecurity, the acknowledgment of the need to unlearn outdated practices, and the promotion of a long-term perspective on continuous learning form the cornerstone of this transformative journey. The roadmap provided in this paper serves as a guide for organisations embarking on this transformative journey, fostering a culture where the ability to adapt is recognised as equally critical as the ability to protect. Ultimately, ACCM represents a paradigm shift, acknowledging that cybersecurity excellence is not a static state but an ongoing, dynamic process of learning, unlearning, and relearning.

Proposed future research directions encompass a range of promising avenues, notably the integration of advanced technologies such as artificial intelligence and machine learning. An emphasis is placed on evaluating the potential of these technologies in automating threat analysis and fortifying defence mechanisms. Additionally, the proposed future research could be around the development of quantitative metrics to measure the effectiveness of adaptive cybersecurity, aiming to furnish an objective evaluation of employee preparedness for paradigm shift.

References

Addae, J. H., Brown, M., Sun, X., Towey, D., & Radenkovic, M. (2017). Measuring attitude towards personal data for adaptive cybersecurity. *Information and Computer Security*, *25*(5), 560–579. https://doi.org/10.1108/ICS-11-2016-0085

Anisetti, M., Ardagna, C., Cremonini, M., Sessa, J., & Costa, L. (2020). Security Threat Landscape. *Security Threats*.

Ask, T. F., Lugo, R., Fritschi, J., Veng, K., Eck, J., Özmen, M.-T., Bärreiter, B., Knox, B. J., & Sütterlin, S. (2023). *The role of cognitive styles and cognitive flexibility in deepfake detection skills and metacognitive accuracy*.

Bada, M., Sasse, A. M., & Nurse, J. R. (2019). Cyber security awareness campaigns: Why do they fail to change behaviour? *arXiv Preprint arXiv:1901.02672*.

Caulkins, B., Marlowe, T., & Reardon, A. (2019). Cybersecurity Skills to Address Today's Threats. In T. Z. Ahram & D. Nicholson (Eds.), *Advances in Human Factors in Cybersecurity* (pp. 187–192). Springer International Publishing.

Craig, J. (2018). Cybersecurity Research—Essential to a Successful Digital Future. *Engineering*, *4*(1), 9–10. https://doi.org/10.1016/j.eng.2018.02.006

Farrelly, J. (2023, October 10). *High-Profile Company Data Breaches 2023*. Electric. https://www.electric.ai/blog/recent-big-company-data-breaches

Gundu, T. (2023a). Chatbots: A Framework for Improving Information Security Behaviours using ChatGPT. In S. Furnell & N. Clarke (Eds.), *Human Aspects of Information Security and Assurance* (pp. 418–431). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-38530-8_33

Gundu, T. (2023b). Internet Of Things: Sensor Layer Security Risk Mitigation Framework. *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1–8. https://doi.org/10.1109/ICECET58911.2023.10389252

Gundu, T., Maronga, M. I., & Boucher, D. (2019). Industry 4.0 Business Perspective: Fostering a Cyber Security Culture in a Culturally Diverse Workplace. *Proceedings of 4th International Conference on The*, *12*, 85–94.

Gundu, T., & Maronga, V. (2019). IoT Security and Privacy: Turning on the Human Firewall in Smart Farming. *Kalpa Publications in Computing*, *12*, 95–104. https://doi.org/10.29007/j2z7

Hielscher, J., Kluge, A., Menges, U., & Sasse, M. A. (2022). "Taking out the Trash": Why Security Behavior Change requires Intentional Forgetting. *Proceedings of the 2021 New Security Paradigms Workshop*, 108–122. https://doi.org/10.1145/3498891.3498902

Jin, C., Xu, A., Zhu, Y., & Li, J. (2023). Technology growth in the digital age: Evidence from China. *Technological Forecasting and Social Change*, *187*, 122221. https://doi.org/10.1016/j.techfore.2022.122221

Kilavo, H., J. Mselle, L., I. Rais, R., & Mrutu, S. I. (2023). Reverse Social Engineering to Counter Social Engineering in Mobile Money Theft: A Tanzanian Context. *Journal of Applied Security Research*, *18*(3), 546–558. https://doi.org/10.1080/19361610.2022.2031702

Madan, S., Savani, K., & Katsikeas, C. S. (2023). Privacy please: Power distance and people's responses to data breaches across countries. *Journal of International Business Studies*, *54*(4), 731–754. https://doi.org/10.1057/s41267-022-00519-5

Mmango, N., & Gundu, T. (2023). Cyber Resilience in the Entrepreneurial Environment: A Framework for Enhancing Cybersecurity Awareness in SMEs. *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1–6. https://doi.org/10.1109/ICECET58911.2023.10389226

Moustafa, A. A., Bello, A., & Maurushat, A. (2021). The Role of User Behaviour in Improving Cyber Security Management. *Frontiers in Psychology*, *12*. https://www.frontiersin.org/articles/10.3389/fpsyg.2021.561011

Netshakhuma, N. S. (2019). Assessment of a South Africa national consultative workshop on the Protection of Personal Information Act (POPIA). *Global Knowledge, Memory and Communication*, *69*(1/2), 58–74. https://doi.org/10.1108/GKMC-02-2019-0026

Quintero-Bonilla, S., & Martín del Rey, A. (2020). A New Proposal on the Advanced Persistent Threat: A Survey. *Applied Sciences*, *10*(11), Article 11. https://doi.org/10.3390/app10113874

Rahman, M. R., Mahdavi-Hezaveh, R., & Williams, L. (2023). What are the attackers doing now? Automating cyber threat intelligence extraction from text on pace with the changing threat landscape: A survey. *ACM Computing Surveys*, *55*(12), 1–36. https://doi.org/10.1145/3571726

Schiller, E., Aidoo, A., Fuhrer, J., Stahl, J., Ziörjen, M., & Stiller, B. (2022). Landscape of IoT security. *Computer Science Review*, *44*, 100467. https://doi.org/10.1016/j.cosrev.2022.100467

Sharma, S., & Lenka, U. (2019). Exploring linkages between unlearning and relearning in organizations. *The Learning Organization*, *26*(5), 500–517. https://doi.org/10.1108/TLO-10-2018-0164

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

Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence*, *1*(12), Article 12. https://doi.org/10.1038/s42256-019-0109-1

Team, I. G. P. (2020). *EU General Data Protection Regulation (GDPR) – An implementation and compliance guide, fourth edition*. IT Governance Ltd.

Thomson, K.-L., & Niekerk, J. van. (2018). Towards Culturally Sensitive Policy: Africanising Approaches to Prevent Social Engineering. *Advanced Science Letters*, *24*(4), 2499–2503. https://doi.org/10.1166/asl.2018.10990

Thomson, K.-L., von Solms, R., & Louw, L. (2006). Cultivating an organizational information security culture. *Computer Fraud & Security*, *2006*(10), 7–11.

Wong, R. Y., Chong, A., & Aspegren, R. C. (2023). Privacy Legislation as Business Risks: How GDPR and CCPA are Represented in Technology Companies' Investment Risk Disclosures. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 82:1-82:26. https://doi.org/10.1145/3579515

Zhao, Y., & Wang, X. (2020). Organisational unlearning, relearning and strategic flexibility: From the perspective of updating routines and knowledge. *Technology Analysis & Strategic Management*, *32*(11), 1251–1263. https://doi.org/10.1080/09537325.2020.1758656

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. https://doi.org/10.1016/j.ijhcs.2019.05.005

Zwilling, M., Klien, G., Lesjak, D., Wiechetek, Ł., Cetin, F., & Basim, H. N. (2022). Cyber security awareness, knowledge and behavior: A comparative study. *Journal of Computer Information Systems*, *62*(1), 82–97.

Anomaly Detection for the MIL-STD-1553B Multiplex Data Bus Using an LSTM Autoencoder

Alec Harlow, Brian Lachine and Vincent Roberge
Royal Military College of Canada, Kingston, Canada

alec.harlow@forces.gc.ca
brian.lachine@rmc.ca
vincent.roberge@rmc.ca

Abstract: Due to the modernization of commercial and military aircraft, real-time systems and their connectivity to ground based networks, including the Internet, that were thought to be "air-gapped", are becoming more susceptible to cyber-attack. Most real-time systems that communicate using the Military Standard 1553B Multiplex data bus (MIL-STD-1553B) protocol do not have the ability to detect cyber-attacks. These systems were originally developed with safety and redundancy in mind, not security. These two factors introduce attack vectors to MIL-STD-1553B communication buses and expose associated avionics systems to exploitation. Recent approaches to anomaly detection for the MIL-STD-1553B data bus have leveraged statistical analysis, Markov Chain modelling, remote terminal fingerprinting and signature-based detection. However, their comparative effectiveness is unknown. Regarding the statistical analysis technique, the lack of accuracy and precision in detecting the start and stop time of anomalous events are not ideal for conducting investigations due to the sheer volume of messages still required to be manually analysed. Deep learning techniques offer an effective means of anomaly detection and applying these techniques to the MIL-STD-1553B data bus could provide more accurate and precise detection times when anomalies or attacks are present, when compared to known statistical analysis, leading to more efficient forensic investigations of anomalous events. The aim of this research is to improve the time-related performance metrics when detecting attacks on the MIL-STD-1553B data bus traffic using a Long Short-Term Memory (LSTM) autoencoder. In order to accomplish this aim, an LSTM autoencoder detector was developed and tested on two separate datasets from different MIL-STD-1553B network architectures, totalling 15 threat instances over 5 scenarios. The detector was then compared to the MIL-STD-1553B Anomaly-Based Intrusion Detection System (MAIDENS) detector, a statistical-based intrusion detection system. The LSTM autoencoder detected every threat instance with no false positive or false negative results and significantly improved the time-related performance metrics when compared to the MAIDENS detector. The results demonstrated this deep learning technique as an effective method for identifying anomalies on a MIL-STD-1553B data bus and significantly reducing the overall number of frames to be analysed during the investigation of identified anomalies.

Keywords: MIL-STD-1553B, Anomaly Detection, Deep Learning, LSTM Autoencoder, Aviation Cybersecurity

1. Introduction

Current MIL-STD-1553B anomaly detection techniques are able to detect a number of attack techniques that include, but are not limited to: Denial of Service (DoS) attacks, Remote Terminal (RT) spoofing attacks, Bus Controller (BC) spoofing attacks and attacks that manipulate RT messages (Stan et al, 2017). The MIL-STD-1553B Anomaly-Based Intrusion Detection System (MAIDENS), a statistics-based intrusion detection system purposed by Généreux et al (2020) is able to detect all of the aforementioned attacks. Other detectors such as a Markov chain model proposed by Stan et al (2017) and a fingerprinting method proposed Stan et al (2019) were able to detect a smaller subset of attack types. The issue faced with a detector like MAIDENS, even though it has an impressive zero false positive rate, is that the detection time accuracy and precision achieved when indicating the start and stop time of an attack is not ideal for conducting forensic analysis on a detected attack. Having a detector that could reduce the detection window and thereby the number of messages associated with potential attack traffic, would significantly improve analysis efforts post attack detection.

To address the need to improve the time-related performance metrics when detecting attacks on the MIL-STD-1553B Data Bus, this paper presents research into the detection of anomalies on the MIL-STD-1553B Data Bus using a Long Short-Term Memory (LSTM) autoencoder deep learning (DL) technique. It identifies the improvement in time-related performance metrics of an LSTM autoencoder detector in identifying attacks in MIL-STD-1553B bus traffic when compared to the MAIDENS detector purposed by Généreux et al (2020).

The statistics-based intrusion detection system, MAIDENS, can detect every attack event or threat occurrence that it has been tested against with perfect event classification accuracy and precision. However, if you consider the time-related performance metrics MAIDENS can only detect an anomaly event within a range of 10 ±8 seconds (Généreux et al, 2020). This equates to about 10,000 ±8000 messages of traffic, based on the data transfer rate from the MAIDENS dataset, and would take an expert a significant amount of time to analyse each message to find the actual start and stop time of an attack. This deficiency stems from the

inability to label each individual message through available MIL-STD-1553B recording method which relies on statistical metrics for event classification instead of the more traditional confusion matrix-based approach, that would identify each message as a true/false positive or negative. Therefore, being able to improve the time-related performance metrics, specifically the detection time accuracy and detection time precision of the start and stop time of an attack, would drastically decrease the number of messages needed to be analysed during an investigation. In this context, accuracy is defined as the closeness of agreement between a test result and the accepted reference value and precision is defined as the agreement between independent test results obtained under stipulated conditions of detection (ISO 5725-1, 1994).

In order to accomplish the aim, the MIL-STD-1553B Data Bus traffic dataset from the work conducted in MAIDENS was used and includes five threat scenarios and baseline traffic. The time related performance metrics of both MAIDENS and the LSTM autoencoder methods to detect these anomalies were then evaluated in terms of improved detection time accuracy and detection time precision.

2. Background

MIL-STD-1553B is a military standard bus communications protocol published in 1973 by the United States Department of Defense (DoD). The standard uses a multipoint topology of remote terminals (RTs) connected by a dual redundancy data bus as depicted in Figure 1. One terminal is designated as the bus controller (BC) and initiates and directs all communication on the bus. A bus monitor (BM) can be attached to the bus, but traditionally fills the role of a data historian, analogous to a “black box” on an aircraft.

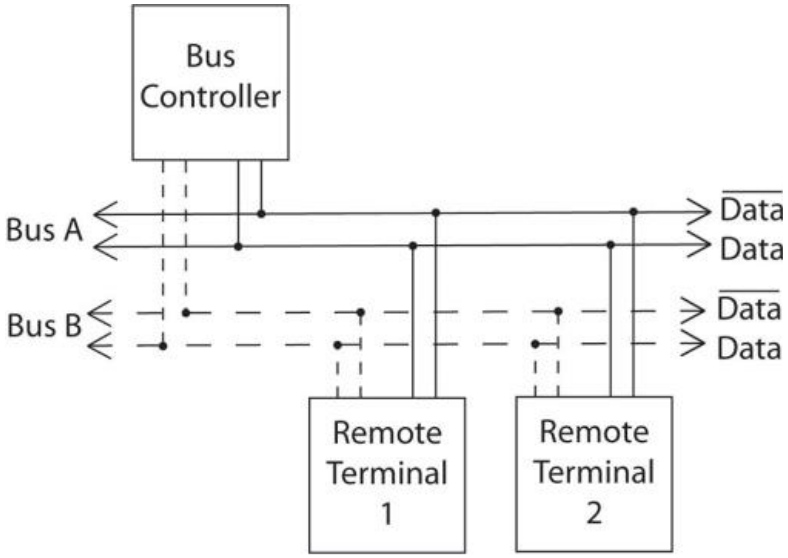


Figure 1: Example of a MIL-STD-1553B Bus Topology Consisting of a BC and Two RTs Connected by a Dual Redundancy Data Bus (Généreux et al, 2020)

There are a total of 32 addresses (0 – 31) on a MIL-STD-1553B bus network, one address (31) is reserved as a broadcast address and the remaining 31 addresses are for potential RTs that can be connected to the data bus. Each address has up to 32 sub-addresses or addressable data buffers that are used to read and write data to and from, with sub-addresses 0 and 31 being reserved for mode codes. Information is transferred through three 20-bit messages: a command word, a status word and a data word. The per-bit-breakdown of each word is shown in Figure 2. The protocol uses these three types of words for all communication on the bus and in order to initiate communication a command word must be sent by the BC (MIL-STD-1553 Designer's Guide, 1998). This ensures that all traffic on the bus is scheduled and reliable.

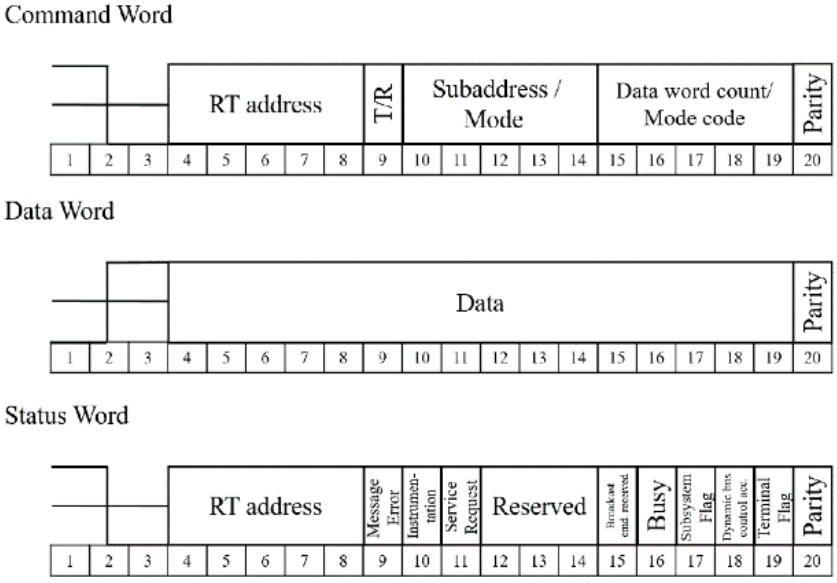


Figure 2: Per Bit Breakdown of Command, Data and Status Words for MIL-STD-1553B (Stan et al, 2019)

2.1 MIL-STD-1553B Vulnerabilities

The MIL-STD-1553B Protocol was designed with safety and reliability, not security in mind. Every communication exchange is preprogrammed and follows a cyclic schedule controlled by the BC. Manufacturers of RTs are expected to follow the guidelines defined by the MIL-STD-1553B protocol (MIL-STD-1553 Designer's Guide, 1998). However, attackers are not limited by these guidelines set by the design documentation and can use its flaws to achieve their desired effect.

Stan et al (2017) showcase two main attack categories: message manipulation and behaviour manipulation. Message manipulation refers to modification of legitimate words (command, data, or status) that are transmitted over the bus. Behaviour modification refers to altering how a component would normally operate. Using these two types of attacks methods, Stan et al (2017) identify 3 main types of threats to the MIL-STD-1553B communications protocol: Denial of Service (DoS), data leakage, and data integrity violation. These attacks can be carried out by two means. The first is by a rogue RT: a device that is not intended to be connected to the data bus and was not part of the original design. The second is by a compromised RT: a RT that is part of the original design, but has been maliciously modified by some means. In addition to the types of attacks described in Stan et al (2017), a study by Lounis et al (2022) reviews and analyses the attack vectors on the MIL-STD-1553B data bus as well. They identify 4 types of attacks that can occur: fabrication attacks, interception attacks, interruption attacks, and modification attacks. These attacks have some overlap with the types of attacks in Stan et al (2017) and contain attack vectors relating to RT components, not just the MIL-STD-1553B protocol. In this paper, the attack taxonomy from Stan et al (2017) will be followed.

2.2 Existing MIL-STD-1553B Detection Methods

The MAIDENS detector purposed by Généreux et al (2020) is a statistical anomaly-based intrusion detection system that uses a time-based, histogram comparison method. The time-based features of the MIL-STD-1553B bus are used to create a baseline histogram representation of known, intrusion-free data. By plotting the frequency of the values of a given feature, they are able to compare run-time data to the baseline data and if it is a certain percentage from the baseline, it is considered anomalous. Based on the results of Généreux et al (2020), the presented MAIDENS detector was able to identify the start and stop times in which a threat occurrence took place across five scenarios with no false positive or negative results. The five scenarios included four DoS type attacks: DoS using Command words, DoS using Status words, RT specific DoS and RT specific sub-address DOS. It also included one data manipulation attack, RT Hijacking, which is a form of data integrity violation and uses a compromised or rogue RT to take over or control a legitimate RT. The detection times ranged from within 0.38 seconds to 47.11 seconds of an attacks start and stop time. This equates to detecting an attack within 380 to 47,110 messages, assuming a bus averages one message every 0.001 seconds, which is not ideal when investigating an attack or threat instance given the potential for a significant number of bus frames to analyse that may not be related to the attack. Overall, the MADIENS detector is

effective at detecting attack threat occurrences, however the accuracy of the detection times could be improved upon to assist in the investigation of detected threat occurrences.

2.3 LSTM Autoencoders

An LSTM autoencoder uses LSTM layers to learn the compressed representations of a dataset (Provotar et al, 2019) and combines the signal reconstruction ability of an autoencoder with the ability of an LSTM to learn and model a time series dataset. The main differences of a LSTM autoencoder and a regular autoencoder is that the main blocks of the network architecture are LSTM cells as shown in Figure 3. The effectiveness of a LSTM in reconstructing the input signal or input values is indicated by its reconstruction error. The higher the reconstruction error value the more inaccurate the predicted values were compared to the input values, the opposite being true for a lower value. This is useful when determining whether data contains anomalies, as the more anomalies contained in the data, the more difficult it will be to recreate the input data resulting in a higher reconstruction error value for each recreated value. The reconstruction error can then be used as a method for setting a threshold value, where any recreated value that is higher than the threshold can be considered anomalous.

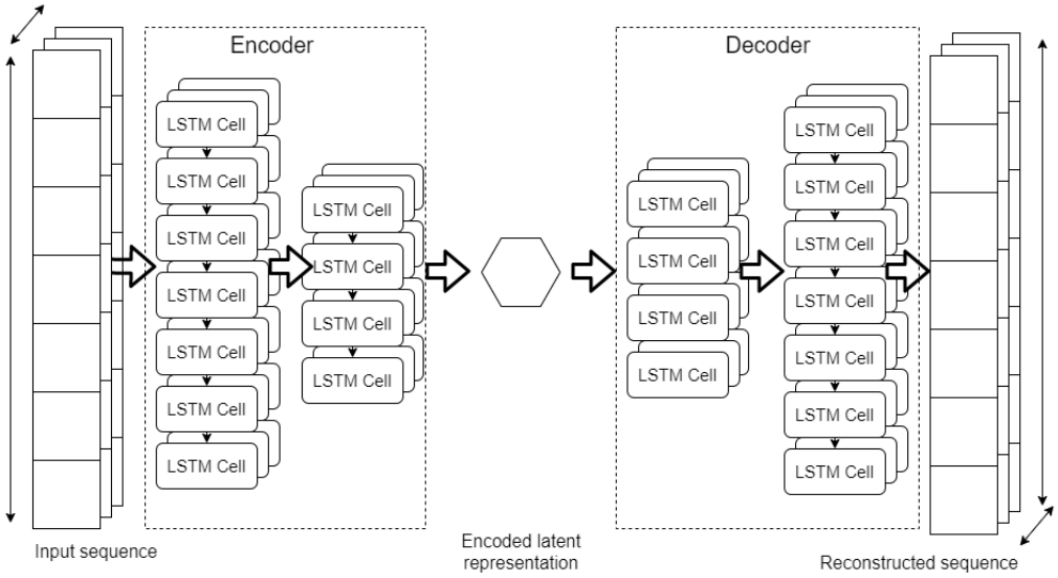


Figure 3: Architecture of a LSTM Autoencoder (Venskus et al, 2021)

2.4 Practical Applications of LSTM Networks

There are examples where LSTM networks either alone or in combination with other methods are used for anomaly detection. One such method was purposed by Taylor et al (2016), using a LSTM RNN to detect anomalies on the automotive Controller Area Network (CAN) bus. The proposed detector accepts raw CAN bus data words as an input and works by learning to predict the next data word originating from each sender on the bus. If there are bits in the next word that are not expected they are flagged as anomalies. This predictive type of modelling would be ideal for behaviour-based anomaly detection, as it is able "to detect data packet anomalies that are unusual only in the context of the rest of the sequence." With the added benefit that, "The LSTM based anomaly detector can be applied to many different vehicles without substantial modification." (Taylor et al, 2016).

As described in Provotar et al (2019), autoencoders that use deep encoders and deep decoders offer many advantages over the standard single layer encoders and single layer decoders. The depth can exponentially reduce the computational cost of representing some functions and also exponentially decrease the amount of training needed to learn some functions. As well, experimentally, deep autoencoders yield a better compression compared to linear autoencoders. In recent years, there has been additional research in the field of using LSTM Autoencoders for anomaly detection. Maleki et al (2021) apply a LSTM autoencoder to detect anomalies in CPU utilization, in a cloud computing environment. They concluded that an advantage of their applied method was the detection of both abrupt and gradually developing anomalies. Said Elsayed et al (2020) use a hybrid method of a LSTM Autoencoder combined with a OCSVM to detect anomalies in web traffic. Their method showed promising results, however due to the nature of web traffic, further real-world

testing was required to confirm the detectors viability. Mahmoud et al (2022) used an LSTM Autoencoder for intrusion detection in IoT systems, which showed impressive performance metrics, when compared to other Autoencoder models on the commonly used NSL-KDD dataset. Overall, the use of LSTM Autoencoder based anomaly detectors have demonstrated the ability to achieve positive anomaly detection results.

The success of the LSTM RNN model in the works of Taylor et al (2019) for anomaly detection on the CAN bus is a good indication that LSTMs will work well for anomaly detection on other bus like networks. Combined with the promising anomaly detection rates of the LSTM autoencoder presented in the works of Provotar et al (2019), Maleki et al (2021), and Mahmoud et al (2022), the MIL-STD-1553B data bus is a good test case for anomaly detection using a LSTM autoencoder DL method, leading the focus of this research to explore the effectiveness of the LSTM autoencoder DL method for anomaly detection on MIL-STD-1553B networks.

3. Methodology and Design

The following three phases were conducted in order to achieve the aim of this research:

1. Data Acquisition Pipeline:
    a. Generation and collection of MIL-STD-1553B baseline and anomalous traffic recordings.
    b. Extraction of features derived from MIL-STD-1553B message traffic.
2. LSTM Autoencoder Model Development:
    a. Creation of a feature extraction tool to ingest MIL-STD-1553B recordings and prepare a usable dataset for model creation.
    b. The creation and application of a LSTM Autoencoder detector to create a baseline model to evaluate subsequent traffic against for anomaly detection.
3. Validation of Detection and Time-Related Performance Metrics
    a. Validation of the detector based upon its ability to accurately and precisely detect anomalies in comparison to MAIDENS using the same datasets.

In order to perform detection on a set of data with a ML or DL technique it has to be in a viable format to be ingested by the detector. The set of data points can be in the form of a stream of data, such as a sequence of messages from a MIL-STD-1553B data bus. In order to collect and create a baseline dataset of MIL-STD 1553B network traffic, a bus recorder was used to record the communications of simulated RTs over a physical MIL-STD 1553B data bus shown in Figure 4. The MIL-STD 1553B data bus was setup similarly to the generic test bench and collection architecture also shown in Figure 4. The data collected was that of a “benign” flight, consisting of an aircraft in flight, where no anomalies were enacted. Once the baseline dataset was collected it was used to create a baseline model in the next phase. The collection of anomalous traffic is similar to the collection of the baseline data except for the presence of a rogue device connected to the physical MIL-STD-1553B bus, to simulate a maliciously attached device that produces anomalies on the bus in the form of different scenarios.

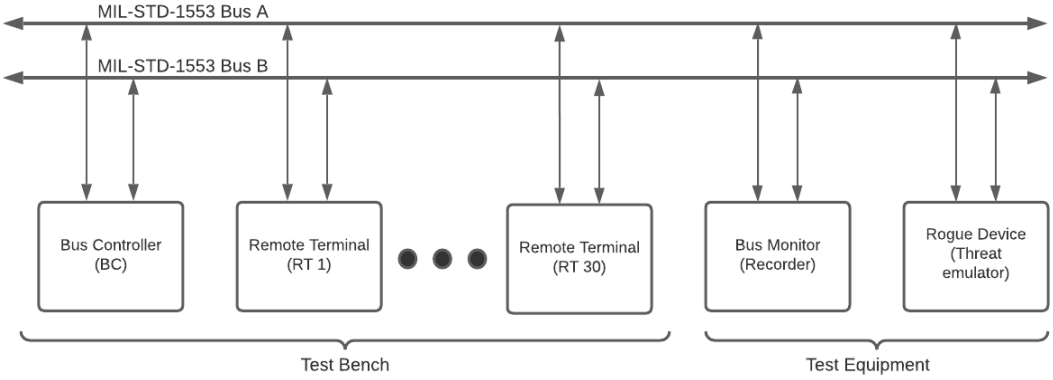


Figure 4: Generic MIL-STD-1553B Test Bench and Collection Architecture

The scenarios were pulled from the following types of attacks:

1. DoS
    a. Network Disruption (Command word DoS)
    b. Network Disruption (Status word DoS)
2. Targeted RT DoS
    a. RT deny
    b. RT subaddress deny
3. Data Integrity Violation
    a. RT hijack

Using the tool presented by Paquet (2014) each of the above five attack scenarios can be carried out on the MIL-STD-1553B bus under test. Each type of attack is described in more detailed in Table 1.

Table 1: Scenario Attack Descriptions

Scenario	Brief Description
1a)	A complete data bus DoS attack where a rogue RT in BC mode uses commands words transmitted on the data bus in order to deny all bus communication.
1b)	A complete data bus DoS attack where a rogue RT in BC mode transmits status words after each normally transmitted command word transmitted on the data bus to deny all bus communication.
2a)	A targeted DoS attack where a rogue RT in BC mode transmits status words after each normally transmitted command word intended for a specified RT to deny communication to or from that RT.
2b)	A targeted DoS attack where a rogue RT in BC mode transmits status words after each normally transmitted command word intended for a specified RT sub-address to deny communication to or from that RT sub-address.
3a)	The implementation of this attack cannot be disclosed due to it being proprietary in nature, however it essentially causes the data for an intended recipient RT to be modified and accepted as if it were the intended values.

The detector was created using the Python programming language with the TensorFlow library as the backend for the LSTM model creation. The LSTM autoencoder model used for this experiment was composed of one input layer, three LSTM layers and one output layer. Sixty-one features were extracted from the MIL-STD-1553B protocol and were used as the inputs for the input layer of the LSTM Autoencoder model. The first LSTM layer was then composed of 30 nodes. The next LSTM layer was then composed of half as many nodes, down to 15 nodes, making the first two LSTM layers, the encoder portion of the LSTM autoencoder network. The last LSTM layer was then expanded back to the size of the first layer, up to 30 nodes, to make the second and third LSTM layers the decoder portion of the LSTM autoencoder network. The final output layer then feeds the results into 61 separate outputs, one for each feature completing the LSTM Autoencoder model.

4. Results

The LSTM autoencoder detector was created using the methodology and design described in the previous section. The detector performs two core functions. The first created a baseline model of the MIL-STD-1553B data bus undergoing analysis. The second function compared subsequent data from the data bus under analysis for anomaly detection. The detector accepts a converted BMDX file (the file format used by Abaco BusTools software) that has undergone feature extraction for baseline creation and anomaly detection in CSV format. The program outputs either a trained LSTM model visualized in Figure 5 or a labelled CSV file that contains all of the messages, the MAE of each message and whether each message was determined to be anomalous based on a set threshold value visualized in Figure 6.

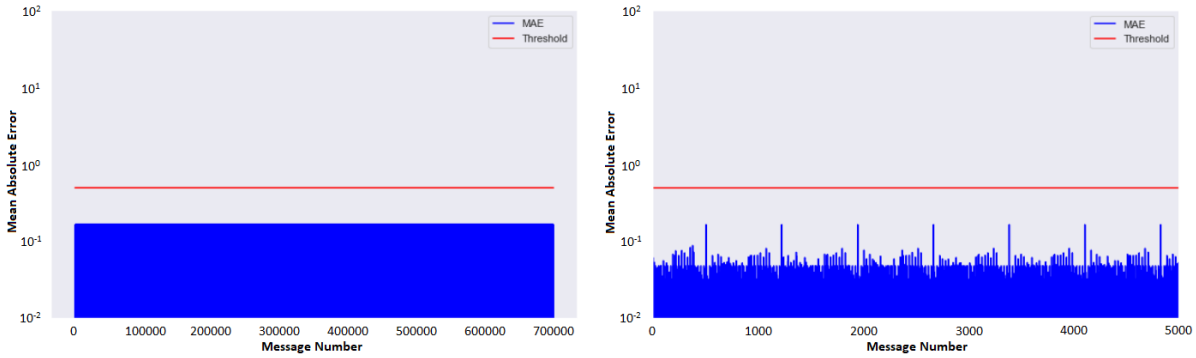


Figure 5: Visualization of Baseline recording (left) same recording scaled down to show fine detail (Right)

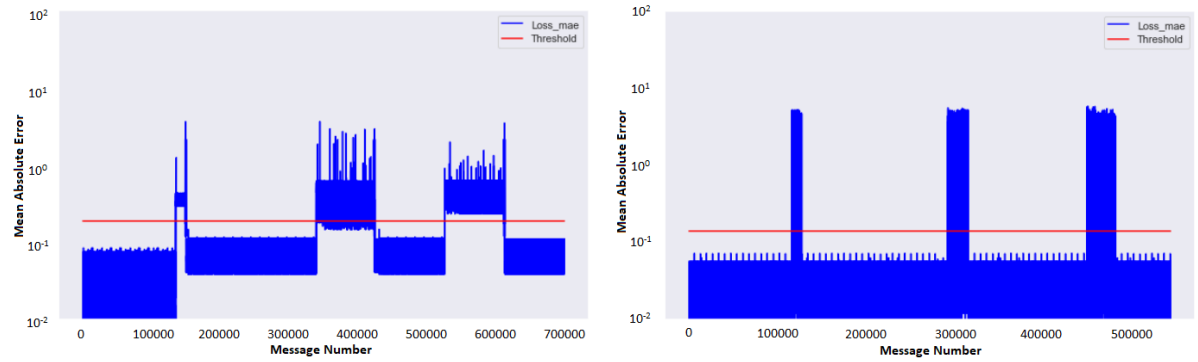


Figure 6: Visualizations of Scenario 1A (left) Scenario 1B (Right)

Figure 5 showcases a baseline LSTM autoencoder model that was created from the dataset. The five scenario recordings that underwent feature extraction were ingested by the model and any of the produced MAE values for each message that fall above the set threshold value were considered to be anomalous. In order to group anomalous messages into events, any messages with a MAE above the threshold were considered a single event once no other messages go above the threshold in a set time interval, in this case a time interval of 30 seconds was used. This allowed for the determination of the start and stop times of the events. Figure 6 showcases the output of the LSTM autoencoder anomaly detector for scenarios 1A and 1B.

All attacks were detected for all five scenarios. When determining only whether an attack has occurred, the event classification accuracy, precision and recall were all calculated to be 100% indicating the LSTM Autoencoder detector had no issues detecting every attack occurrence. Table 2 shows the average detection time accuracy and detection time precision results for each scenario as well as a combined average from all scenarios. The LSTM autoencoder detector detected the start and stop time of the first scenario within 0.0060 ±0.0058 and 0.0060 ±0.0031 seconds, respectively. This means that the detector was within 6 ±6 and 6 ±3 messages of determining the start and stop time of the threat scenarios, respectively. The overall average start and stop time for all combined scenarios is 0.0512 ±0.0354 and 0.2476 ±0.3194 seconds for determining an anomalous event or 50 ±35 and 250 ±320 messages.

Table 2: LSTM Autoencoder Detection Time Accuracy and Precision Results

Scenario	Average Start Time Difference (seconds)	Average End Time Difference (seconds)
1a)	0.0060 ±0.0058	0.0060 ±0.0031
1b)	0.0033 ±0.0011	0.0060 ±0.0075
2a)	0.0045 ±0.0004	0.0039 ±0.0007
2b)	0.04791 ±0.0378	1.1964 ±1.5465
3a)	0.1943 ±0.1322	0.0587 ±0.0391
Average	**0.0512 ±0.0354**	**0.2476 ±0.3194**

The MAIDENS detector used the baseline BMDX file that was converted into CSV format to create a baseline profile. The baseline profile was then used determine the threshold used to detect anomalies in subsequent bus traffic with attack data that was analysed by the MAIDENS detector. For this research all of the MADIENS data was re-run through the detector and the same results were obtained as in the works of Généreux et al (2020). When determining only whether an attack has occurred, the event classification accuracy, precision and recall were all calculated to be 100% indicating the MAIDENS detector also had no issues detecting every attack occurrence. Table 3, shows the overall detection time accuracy and detection time precision results for each scenario. The MAIDENS detector, detected the start and stop time of the first scenario within 0.38 ±0.15 and 47.11 ±37.53 seconds, respectively. This means that the detector was within hundreds to thousands of messages of determining the start and stop time of the threat scenarios. The average start and stop time for all combined scenarios is an average of 0.64 ±0.40 seconds for the start time and 10.27 ±7.98 seconds for the stop time.

Table 3: MAIDENS Detection Time Accuracy and Precision Results

Scenario	Average Identified Start Time Difference (seconds)	Average Identified End Time Difference (seconds)
1a)	0.38 ±0.15	47.11 ±37.53
1b)	0.45 ±0.26	0.89 ±0.13
2a)	0.41 ±0.46	0.78 ±0.32
2b)	0.81 ±0.57	2.26 ±1.77
3a)	1.14 ±0.54	0.3 ±0.16
Average	**0.64 ±0.40**	**10.27 ±7.98**

The LSTM autoencoder anomaly detector outperformed the MAIDENS detector in every scenario. Derived from the data in Table 4, there was an 1054.69% increase in the average detection start time and a 4147.82% increase in the average detection end time when comparing detection time in seconds. In terms of messages, the MAIDENS detector can detect the start time within 640 ±400 messages and the stop time within 10270 ±7980 messages, whereas the LSTM autoencoder detector can detect the start time within 51 ±35 messages and the stop time within 250 ± 320 messages. Based on the median overall average, the LSTM autoencoder anomaly detector is 1058.82% more effective at detecting the start time and 4108% more effective at detecting the end time of an anomalous event in terms of messages than the MAIDENS detector.

Table 4: Average Start and End Detection Times in Seconds

	Average Start and End Detection Time (Seconds)					
	Start Low	Middle	Start High	End Low	Middle	End High
LSTM-Autoencoder	0.0158	0.0512	0.0866	-0.0718	0.2476	0.567
MAIDENS	0.14	0.54	0.94	2.29	10.27	18.25
% Increase	886.076	1054.69	1085.45	3189.42	4147.82	3218.70

There were a couple of drawbacks that were noted during experimentation. The first is the manual setting of a threshold value. For both datasets an acceptable threshold value was able to be set for all anomaly detection. However, depending on how the MIL-STD-1553B protocol was implemented or what state of flight an aircraft is in during recording, this value will likely need to be adjusted accordingly. The setting of the threshold value may require expert knowledge on the system and is a significant factor to ensure the model is detecting only anomalous events. The second drawback noted related to data collection and how the threat emulation tool from Paquet (2014) did not have the ability to tag individual messages as anomalous. The LSTM autoencoder by design is able to tag each message as anomalous or not. Overall, the increase in performance when compared to MADIENS is worth the few drawbacks, with each drawback that could be studied and expanded on further in the future to potentially minimize their effect.

The above results showcase the viability of DL methods as a means of detecting of anomaly on a MIL-STD-1553B Bus. An LSTM autoencoder is just one example of many available DL methods that can be explored for anomaly detection on network architectures like the MIL-STD-1553B protocol. Advances in means of detection

of anomalies can not only inform users of intrusions on their networks, but can also be a means of fault detection for maintenance actions.

5. Conclusion

While the detection of an actual cyber-attack on a physical aircraft has yet to be publicly realized, the consequences of such an event cannot be ignored. There are methods that have been developed for the MIL-STD-1553B protocol and include signature-based, and anomaly-based detection techniques. The aim of this research was to improve the time-related performance metrics when detecting attacks on the MIL-STD-1553B data bus traffic using a LSTM autoencoder deep learning technique.

Not only was it successful in detecting attacks in five different scenarios, but it also significantly outperformed the MAIDENS detector presented by Généreux et al (2020) in both detection time accuracy and detection time precision. The LSTM autoencoder detector indicated the start and stop timings of each anomalous scenario to a hundredth or thousandth of a second with no false positive or negative results. The degree of detection accuracy would be more beneficial, compared to MAIDENS, in a forensics investigation as the detected time are closer to the actual attack time, resulting in far fewer frames to needlessly analyse. This work provides substantial evidence of the applicability of the LSTM autoencoder detector for anomaly detection on MIL-STD-1553B networks.

References

Généreux, S. J. J., Lai A. K. H., Fowles C. O., Roberge, V. R., Vigeant, G. P. M. and Paquet, J. R. (2020) "MAIDENS: MIL-STD-1553 Anomaly-Based Intrusion Detection System Using Time-Based Histogram Comparison", *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 56, No. 1, February, pp. 276-284.

ISO 5725-1 (1994), "Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions", [online], https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en

Lounis K., Mansour Z., Wrana M., Elsayed M. A., Ding S. H. H., and Zulkernine M. (2022) "A Review and Analysis of Attack Vectors on MIL-STD-1553 Communication Bus", *IEEE Trans. Aerosp. Electron. Syst.*, pp. 1–1.

Mahmoud M., Kasem M., Abdallah A. and Kang H. S. (2022) "AE-LSTM: Autoencoder with LSTM-Based Intrusion Detection in IoT", *2022 International Telecommunications Conference (ITC-Egypt)*, Alexandria, Egypt, pp. 1-6.

Maleki S., Maleki S., and Jennings N. R. (2021) "Unsupervised anomaly detection with LSTM autoencoders using statistical data-filtering", *Appl. Soft Comput.*, Vol. 108, p. 107443, September.

"MIL-STD-1553 Designer's Guide" (1998), ILC Data Device Corporation.

Paquet J. (2014) "Uncovering MIL-STD-1553 vulnerabilities: exploitability of military aircraft networks", MASc, Royal Military College of Canada.

Provotar O. I., Linder Y. M., and Veres M. M. (2019) "Unsupervised Anomaly Detection in Time Series Using LSTM-Based Autoencoders", *IEEE International Conference on Advanced Trends in Information Theory (ATIT)*, December, pp. 513–517.

Said Elsayed M., Le-Khac N. A., Dev S., and Jurcut A. D. (2020) "Network Anomaly Detection Using LSTM Based Autoencoder", *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, Alicante Spain, November, pp. 37–45.

Stan O., Elovici Y., Shabtai A., Shugol G., Tikochinski R., and Kur S. (2017) "Protecting Military Avionics Platforms from Attacks on MIL-STD-1553 Communication Bus", *ArXiv170705032 Cs*, July, [online], http://arxiv.org/abs/1707.05032

Stan O., Cohen A., Elovici Y., and Shabtai A. (2019) "Intrusion Detection System for the MIL-STD-1553 Communication Bus", *IEEE Trans. Aerosp. Electron. Syst.*, pp. 1–1, 2.

Taylor A., Leblanc S., and Japkowicz N. (2016) "Anomaly Detection in Automobile Control Network Data with Long Short-Term Memory Networks", *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, October, pp. 130–139.

Venskus J., Treigys P., and Markevičiūtė J. (2021) "Unsupervised marine vessel trajectory prediction using LSTM network and wild bootstrapping techniques", *Nonlinear Anal. Model. Control*, Vol. 26, No. 4, pp. 718–737.

Authentication in a Hyperconnected World: Challenges, Opportunities and Approaches

Christoph Lipps[1], Jan Herbst[1], Rekha Reddy[1], Matthias Rüb[1], and Hans Dieter Schotten[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]University of Kaiserslautern (RPTU), Germany

Christoph.Lipps@dfki.de
Jan.Herbst@dfki.de
Rekha.Reddy@dfki.de
Matthias.Rueb@dfki.de
Hans_Dieter.Schotten@dfki.de

Abstract: Authentication and integrity are the prerequisites for trustworthy and secure communication. Without unambiguous knowledge of who is being interacted with, no confidential content can be exchanged, no (remote) access to systems and equipment can be granted, and no trust can be established. This situation is further exacerbated by an increasing interconnection and globalization towards a hyperconnected world. (Communication) Participants are no longer necessarily in close physical and social proximity and do not need to know each other, but can have their source/destination anywhere in the world. An authentication process is used to verify that someone -whether human or machine-, is in fact who she claims to be: The process thus includes a validation step to evaluate an assertion. However, systems differ in terms of their requirements, for instance with regard to the authentication options available, the time period required for re-authentication and the frequency of re-authentication, as well as the level of security to be achieved with authentication. The latter particularly with regard to the cost/benefit ratio of the application. Additionally, there are efforts to finally abolish traditional passwords, passphrases and pin codes and render them obsolete. In this work, technologies and methods for authentication beyond passwords and trustworthy authentication will be examined, particularly with regard to future communication infrastructures such as Beyond 5G and Sixth Generation (6G) wireless systems. Thereby, the impact of Artificial Intelligence (AI) methods, but also the relevance to Quantum Key Distribution (QKD) and Post Quantum Cryptography, as well as the use of 6G-enabling technologies like Reconfigurable Intelligent Surfaces (RISs), Wireless Optical Communication (WOC) and Physical Layer Security (PhySec), for example as additional factors of a multi-factor authentication process, will be considered, along with Body Area Networks (BANs) and the integration of the human body relying on biometrics. The various concepts are compared with regard to their requirements, limitations and possible applications in order to provide the user with an orientation as to which authentication method is conceivable and useful in which specific scenarios.

Keywords: Beyond 5G, 6G, Physical layer security, Physically unclonable functions, Human-PUFs, Channel-PUFs, (Cyber)Security, Sixth generation

1. The Hyperconnected World: A Vision for the Future

The future will be inter- and hyperconnected: According to the Cisco Annual Internet Report, 2018 there were around 480 million interconnected devices, by 2023 there are 823 million, however, it is expected that by 2030 there will be more than 50 billion interconnected devices with a data volume transmitted of up to 350 to 400 zetabytes (Cisco, 2020). Communication systems will thereby be the core element of digital value creation and society, as well as the technological foundation for digitalisation and thus the enabler of innovation and progress. Particularly in sectors such as the (Industrial) Internet of Things (I)IoT, right the way through to an industrial metaverse in which machines, factories, buildings, cities, networks and transport systems are virtually mirrored and digitally twinned (DT), via applications such as Augmented and Extender reality (AR/XR), and autonomous driving, big data analytics and global connectivity will be commonplace. As highlighted in **Figure 1**, a number of additional factors contribute to the hyperconnectivity scenario. These include the integration of low orbit satellite communication, drones and space exploration technologies, as well as aspects of sustainability. In this all-electric society with smart cities and connected healthcare system - up to a medical metaverse -, the confidentiality, reliability and resilience of networks and systems are essential. Artificial Intelligence (AI) methods will be integrated at every level of the processing chain, in central nodes and cloud environments, to calculate and train complex systems, but also in the fog and at the edge (Kimovski, et al., 2021). This requires an appropriate infrastructure in terms of bandwidth, latency and throughput on the one side, and on the other side there are demands on the reliability of the AI in terms of Trusted AI and Trustworthy AI (Ma, et al., 2023).

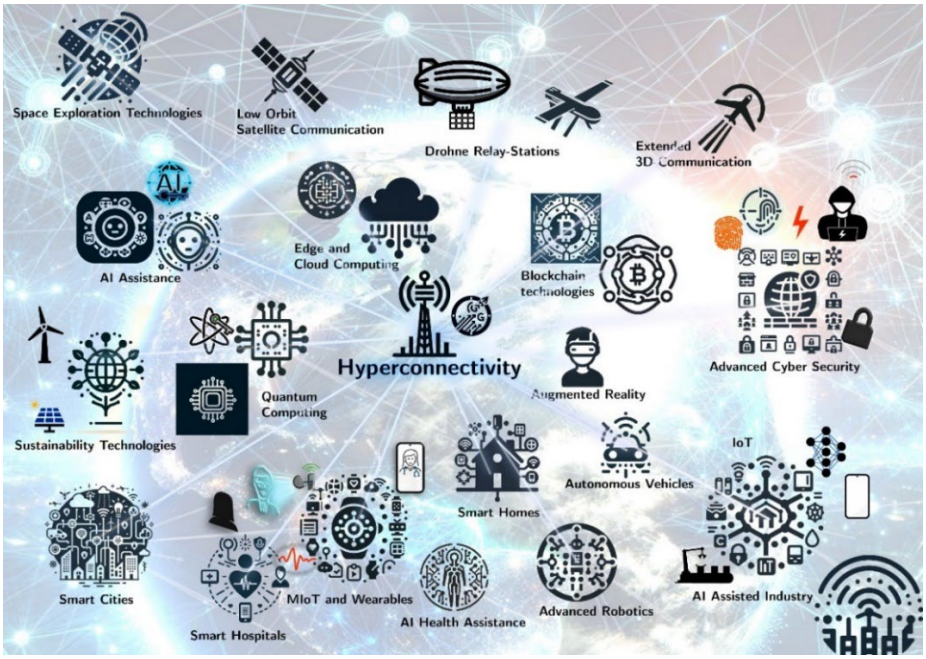


Figure 1: Hyperconnectivity in the future: efficient, sustainable and secure hyperconnectivity as an enabler of an all-connected world[1]

Although the quest for security and trust is perhaps as old as humanity itself -potential threats posed by strangers have always had to be recognised, mitigated and at best prevented-, trust must be developed, maintained, and reinforced (Lipps, et al., 2023). But with increasing globalisation and interconnection, combined with the ability to remotely access systems -it is no longer necessary to be in close proximity to access a system-, and with the rising importance of transmitted data in terms of confidentiality and privacy, the motivation for hackers and saboteurs is rising as well. For instance, the attacks to Cyber-Physical critical infrastructure, such as to water systems (Tuptuk, et al., 2021), the Colonial Oil Pipeline (Lipps, et al., 2022) and the KillNet Distributed Denial of Service (DDoS) attack to the health and energy sector (Tolanur & Chaudhari, 2023) are evolving and are every day threats (Ciocarlie & Zhou, 2023). Furthermore, there will be a series of AI-based attacks in the future. These include, among others, attacks with the assistance of AI, as it is already happening today with deepfakes (Chadha, et al., 2021), for example; but also attacks on AI, such as ii) the conscious manipulation of data by data poisoning (Tolpegin, et al., 2020) or backdoor attacks on AI models (Lin, et al., 2020).

The hyperconnected world of tomorrow is promising many simplifications, improvements and new capabilities, but its realisation will significantly depend on the success of a series of enablers:

- Powerful, efficient and sustainable telecommunication infrastructure => Beyond 5G and 6G technology
- Artificial Intelligence and their applications => Distributed AI, Trusted AI and Trustworthy AI
- Cyber-Resilience and Cyber Security => Security, Privacy and Resilience of infrastructure and applications

To elaborate on this, the rest of the work is structured as follows: A definition of the concepts of "identification" and "authentication", as well as a categorisation and the current status, are given in Section 2. Furthermore, Section 3 addresses the challenges arising and proposes technological possibilities for future methods. These are discussed in Section 4 and an overview of the strengths and weaknesses is given, followed by Section 5, which concludes the work and provides an outlook for future work.

2. Identification and Authentication: The Status Quo

In a hyper-networked world, as in the field of information security and social relations in general, the functions of identification and authentication are of fundamental importance and are strictly related. Whereas

[1] Please note: This image was generated with the assistance of AI

identification is the process by which an entity claims its identity, authentication is the proof, the validation of this claim. These concepts are emphasised in more detail and brought together for secure communication in the future.

2.1 Understanding Identification and Authentication

Especially because (biometric) identification is as old as humanity itself (see Biometrics) and there is a multitude of more or less sophisticated methods for validation available over time, a basic definition and objective understanding of the term is necessary. Therefore, the National Institute of Standards and Technology (NIST) describes authentication as “a process of establishing confidence” (Burr, et al., 2013). In combination with *Burrows, Abadi & Needham* which are specifying it in detail with the technical perspective, that "after authentication, two principals should be entitled to believe that they are communicating with each other and not with intruders" *(1990)* and together with *O’Gorman*’s requirements about a “process of positively verifying the identity” *(2003)* of an entity, an adequate indication can be made: *establishing confidence* and *verifying the identity* form the very fundamentals of a trustworthy communication whereupon secure exchange of information can be built on.

Already with the emerging idea of distributed systems and the growth of network technology in the 1990s, *Burrow, Abadi & Needham* demanded that "authentication protocols are the basis of security in many distributed systems" *(1990)*. This is emphasised by *Bruce Schneier* in relation to access control, which cannot be carried out until the identification -"who are you"-, and authentication -"and can you prove it"-, have been completed (Schneier, 2000). Historically, several protocols have decisively influenced the modern understanding of authentication. These include the Needham-Schröder protocol (Needham & Schroeder, 1978), Otway-Rees (Otway & Rees, 1987) and Kerberos (Garman, 20023), each contributing to the development of secure authentication mechanisms in various systems. Thereby, the practical aspect of authentication involves a dual responsibility, granting access to legitimate users, while at the same time denying and limiting it to unauthorized entities.

Furthermore, *Schneier* insightfully links the concepts of security and trust, by highlighting that “security is what you need when you don’t have any trust” and “security is ultimately […] (to) introduce trust in society” (Schneier, 2012). However, at the same time, he also emphasises the holistic approach to authentication and security as “cryptography is not a panacea – you need a lot more that cryptography to have security – but it is essential” (Schneier, 2000). The factors influencing the effectiveness of identification and authentication process include the availability of information, the type of the system and use-case, the access level and the potential use of cryptographic keys.

2.2 Multi-Faktor Authentication: Principle and Factors

In the heterogeneous landscape of information security, the development of authentication methods has been driven primarily by the balance between robust security and user-friendly access. Over decades, passwords, for instance, have been the cornerstone of secure systems. Their simplicity and ease of implementation made them ubiquitous in almost every application. However, passwords also have significant weaknesses, as Bruce Schneier emphasizes: For convenience, users often choose a password to easy to guess, whereas if they are forced to generate a complex one, they write it on a post-it (Schneier, 2000). The habit of using the same password for multiple platforms, as well as the willingness to share a password with others, especially when it is helping to get work done, exacerbates this vulnerability (Schneier, 2000).

Given these limitations, there are efforts in academia to replace text passwords. In their comprehensive study *Bonneau, et al.* (2012) describe these efforts and discuss different approaches, such as graphical and cognitive methods, which they contrast in detail. In addition, there have been approaches for many years, for example to use existing social media accounts (social login) (Ho & Katuk, 2016) or to be able to access multiple applications with a single set of credentials (single sign-on (SSO)) (Radha & Reddy, 2012).

If a single factor, such as a password, is used, this is referred to as Single-Factor Authentication (SFA). Adding another factor to the system is called 2-Factor Authentication (2FA), while 3-Factor Authentication is called Multi-Faktor Authentication (MFA) and 4-Factor Authentication or more is known as Strong Authentication (SA) (Schneier, 2015). Typically, MFA is designed to use factors from different areas: i) something you know (such as passwords); ii) something you have (cards or tokens); iii) something you are (biometrics). There are also current efforts to include context as an additional factor in the process to increase plausibility and security (Huseynov & Seigneur, 2017).

Nevertheless, on the path to replacing conventional mechanisms, it will play a decisive role how the methods of artificial intelligence and the distributed infrastructure will be treated. In the future, it will be more important to operate and learn on encrypted data (Smart, 2023) and to use methods such as Fully Homomorphic Encryption (FHE) (Sanon, et al., 2023) in order to take privacy aspects into account, as well.

3. Authentication of Tomorrow: Challenges and Approaches

With the ongoing interconnection and the increasing use of distributed applications, the available authentication options are also changing significantly. Here, the emerging challenges but also innovative approaches in the field of authentication are addressed, especially with regard to the wireless communication of the future, in relation to Beyond 5G (B5G) and the Sixth Generation (6G) (Ruzomberka, et al., 2023) wireless technologies. Machine Learning (ML) approaches as well as Zero Trust Architectures (ZTA) and zero knowledge authentication will also be considered.

3.1 Biometrics: A Step Towards Personalized Security

Biometric authentication is the oldest existing method, but it is not just a human characteristic, even animals can recognize their counterparts by smell, stature and voice (Schneier, 2014). In the meantime, however, this area has developed beyond the traditional methods such as voice, fingerprint or iris pattern and more sophisticated methods are being used. These include the use of bio-electrical signals emitted by the human brain in the form of Electroencephalography (EEG) (Bidgoly, et al., 2020), the volumetric changes of peripheral blood vessels via Photoplethymography (PPG) (Sarkar, et al., 2016) or the use of step and movement patterns recorded by intelligent insoles (Lipps, et al., 2021). While biometrics, which enable the authentication of individuals based on unique personal characteristics, have been acclaimed as a significant advance in security technology, and fundamentally "biometrics are great because they are really hard to forge" (Schneier, 2000), they also suffer from inherent difficulties as they are "lousy [...] (and) easy to forge" (Schneier, 2000). Above all, the emergence of highly developed AI and Deepfake (Chadha, et al., 2021) technologies fostered new vulnerabilities. The ability to imitate voices, facial features and even fingerprints raises serious security concerns and renders traditional biometric methods practically ineffective. In addition, the problem remains that once biometric data has been compromised, it can no longer be used as a feature.

3.2 Physical Layer Security

According to *Lipps and Schotten*, Physical Layer Security (PhySec) comprises “various methods of how to utilize inherently present characteristics of media (hardware, wireless channel and human body) to derive secrets applicable as cryptographic primitives” *(2022)*. Besides "traditional" methods such as Physically Unclonable Functions (PUFs) (Gassend, et al., 2002) -using properties of semiconductor devices and electronic circuits-, and Secret Key Generation (SKG) -using properties of the wireless radio channel-, these also include Human PUFs -properties of the human body; biometrics-. In particular, the so-called 6G-enabling technologies such as Wireless Optical Communication (Wei, et al., 2022), Reconfigurable Intelligent Surfaces (Lipps, et al., 2023) and AI, open up a wide range of possibilities for incorporating these methods into the security and authentication process.

Following *Bruce Schneier's* statement "security is never black and white, and context matters more then technology" (Schneier, 2000), the effectiveness and appropriateness of security measures depends significantly on the context in which they are applied. Different technologies, each with their own strengths and weaknesses, need to be considered as part of a comprehensive security strategy. This approach is crucial for developing systems that are not only secure, but also practical and user-friendly. Precisely the PhySec methods can contribute to this, not only depending on the context, but also as an additional factor, as additional context in MFA applications: If, for example, the MFA can be complemented by characteristics of the wireless channel, which are almost unique in this form, or if a RIS can provide such individual characteristics, then this is a major step towards more secure, plausible and integer authentication.

3.3 Zero Trust Architectures and Zero Knowledge Authentication

The principle of Zero Trust Architecture (ZTA) is becoming increasingly important, especially in the context of the security of the next generation of mobile communications, 6G. *Chen*, for example, emphasizes the importance of ZTA in establishing secure network environments (2023). This approach is based on the assumption that no entity, whether internal or external, is inherently trustworthy - never trust, always verify - even if the entity is already within a trustworthy environment. Authentication is therefore required at every stage of digital interaction.

As the principle of zero trust can be applied to data access and data management, the implementation of privacy-preserving authentication techniques requiring authentication every time data is accessed represents a significant step forward in protecting user data and ensuring secure access (Tang.2023). Especially with regard to corresponding user roles, such as Attribute-Based Access Control (ABAC), in combination with context-based authentication (see above), there are opportunities for innovative solutions.

In addition to ZTA and the basic distrust of entities and access, authentication without a priori knowledge, Zero Knowledge Authentication (ZKA), is a pivotal technology to improve privacy and security, especially in (I)IoT and B5G/6G networks of the future. *Chen's* study highlights the potential of ZKA in digital authentication, which allows credentials to be verified without revealing sensitive information beforehand (2023). Even though the basic principle dates back to the early 1990s (Bellovon & Merritt, 1992), the ZKA addresses the challenges of (I)IoT by authenticating devices and users without compromising their sensitive data. This approach is crucial in networks with extensive data collection and transmission. In particular, due to its efficiency and scalability, ZKA offers advantages that fulfil the requirements of high-performance and densely interconnected environments.

Furthermore, the combination of ZKA with ML (Arp, et al., 2023) increases security by detecting unauthorized access and anomalous behaviour, further enhancing the security of IoT networks. The use of ZKA in IoT and advanced wireless networks is a significant step towards a secure digital ecosystem and meets the growing demand for privacy-friendly and user-friendly authentication solutions.

4. Secure Authentication: An Evaluation

In the dynamic field with a wide range of different authentication methods, the evaluation of security and efficiency is of crucial importance. Based on the previous descriptions, corresponding criteria for a rather objective evaluation of methods are proposed here. Therefore, based on the proposal by *Bonneau, et al.* (2012), a distinction between *usability*, *deployability* and *security,* the principle is extended at this point to encompass the dimensions *continuity* and *adaptability*[2]. As illustrated in Figure 2, authentication primarily needs to be thought of as layers. All domains are important, no doubt, but depending on the use case and available resources, the requirements have to be prioritized as suggested here: The Security & Resilience of the applications ought to be at the core of the requirements. As listed in Table 1, each of these categories consists of a set of relevant requirements. For instance, components such as resilience against leakage, phishing or the guessing of factors, but also cryptographic strength and unlinkability are significant for the Security & Resilience of a system.

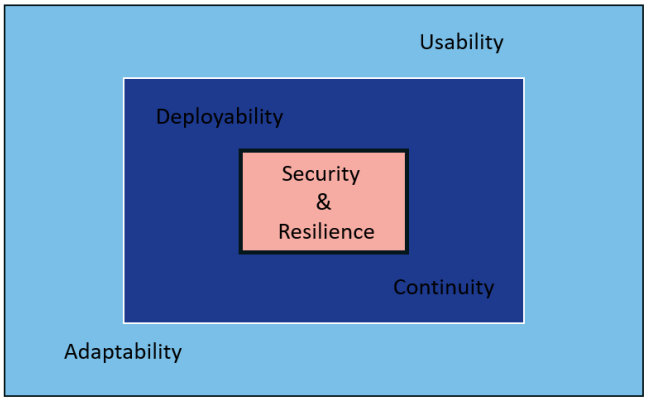


Figure 2: Layered model of the requirements for authentication mechanisms

In a second layer, the requirements for the deployability and continuity of a method are to be considered. This includes, for example, how practical and feasible a method is, but also aspects such as the validity period of a completed authentication or the ability to re-authenticate. The aspects of usability and adaptability can be considered in the outer layer, and thus as merely subordinate and decisive. An overview of the significance of the factors and details regarding their content is provided in Table 1.

[2] Please note: The ranking of components given here is the authors' suggestion; there may be other, individual rankings depending on use cases and available resources;

Table 1: Overview and description of the criteria

Criteria	Description
Security & Resilience	How secure and resilient is a method/system against attacks and intentional and unintentional incidents?
Resilience against compromising (spying, eavesdropping, faking, phishing,…)	Resistance to the loss of a factor through espionage, wiretapping or faking
Resilience against theft	Resistance to compromise through theft of single or multiple authentication factors.
Resilience against guessing and imitation	Resistance to randomness (entropy) and deception
Cryptographic strength	Strength of a method in terms of cryptographic properties such as complexity, computing power and memory
Unlinkability	Inability to determine whether users authenticate themselves multiple times; ability to keep different sessions separate and not linked
Deployablity	How operational is a method in terms of its feasibility and practicability of implementation?
Accessibility	Accessibility for users with different abilities and requirements; loss/restriction of physiological abilities
Maturity of the application	Technical and practical readiness of a system; functional implementation available?
Proprietary application?	Is the method a proprietary system?
Costs caused by the application (time, money, computing power,...)	How expensive is the application in terms of computing power, monetary resources, time for implementation and execution
Continuity	How continuous is the application, particularly in terms of period of validity and combinability
Validity period of authentication	How long is an authentication valid, in which time frame is a new authentication necessary?
Re-authentication capability	Does the method enable re-authentication? How complex/expensive is it?
Combinability with other methods	Can the method be combined with other methods (also in the context of an MFA)?
Usability	How manageable is a method?
Complexity of the method (ability to learn it, complexity of the steps to be completed,...)	Does the method require a complex and elaborate procedure or can it be used by everyone?
Need for additional hardware/tokens/aids	Is anything additional required to use the method, such as hardware or other specialized information?
Adaptability	How adaptable is the method to internal and external influences and current events?
Scalability of the method	How many devices can the method operate, is there an upper limit?
Adaptability to internal and external changes	Can the method react or be adapted to unforeseeable internal/external influences?

In addition to the criteria listed in **Table 1**, there may be others; the list is deliberately designed in such a way that additions can be made at any time. The selection of the superior groups is also not a closed list, although an attempt has been made to define these categories as universally as possible. Current situations and new technological developments lead to further adaptation and integration of the list, this is possible at any time. However, this overview can serve as a starting point, especially with regard to future work. **Table 2** correlates the above criteria with existing authentication methods. The methods were discussed and classified by a panel of experts.

Table 2: Evaluation of the criteria with existing authentication methods

		Passwords, PIN-Codes, Passphrases	Biometics (physiological, behavioural and. biological)	Certificates	Token, Smart-Cards	Trusted Platform Modules, Hardware Security Modules, Physically Unclonable	Zero-Trust Architectures	Zero-Knowledge Authentication	Context (location, date, plausibility, Physical Layer Security Factors)
Security & Resilience	Resilience against compromising	o[3]	✓[3]	o	✓[3]	✓	✓	✓	o
	Resilience against theft	X	✓[3]	X	X	o	✓	✓	✓
	Resilience against guessing and imitation	o[3]	✓[3]	✓	✓	✓	✓	✓	✓
	Cryptographic strength	o[3]	o[3]	✓	✓	✓	-	-	-
	Unlinkability	X	✓	✓	o	✓	-	-	-
Deploy-ability	Accessibility	✓	o	✓	✓	✓	✓	✓	✓
	Maturity of the application	✓	o	✓	✓	✓	o	o	o
	Proprietary application?	X	X	X	X	X	X	X	X
	Costs caused by the application	L	H	M	M	M	M	M	M
Continuity	Validity period of authentication	o	o	L	L	L	-	-	M
	Re-authentication capability	H	H	M	M	H	-	-	H
	Combinability with other methods	H	H	M	H	H	-	-	H
Usability	Complexity of the method	L	M	M	M	H	H	H	M
	Need for additional hardware/tokens/aids	X	X	✓	✓	✓	-	-	✓
Adaptability	Scalability of the method	H	M	H	M	M	H	H	H
	Adaptability to internal and external changes	X	o	X	X	X	-	-	✓

✓= Yes, X = No, o = Medium, M = Medium, L = Low, H = High, - = no consideration

As mentioned, the summary in **Table 2** provides an overview of the different categories of existing authentication groups and the criteria for their use. There is no claim to completeness here, but should serve as a starting point for a deeper analysis of the subject area. The different methods are not always consistent and depend in part on the specific application and method. For example, "Resilience against imitation" is marked as good for biometric methods, although it is well known that there are attacks by deepfakes, for example, or in the knowledge that fingerprints and iris scans have been considered broken for years. However, modern and sophisticated procedures such as gait recognition or EEG are taken into account.

5. Conclusion and Future Work

The future of all of us will be a hyperconnected future. Above all, we will be able to perceive the effects clearly in everyday applications such as autonomous driving, medical applications and living in the metaverse. But as not only the amount of data will increase significantly, but its content and value will gain in importance, its integrity and security will be more crucial than ever before. The insight given into the multifaceted field of authentication highlights that the evolution towards secure, efficient and user-friendly authentication is an ongoing and evolving challenge. Technological advances, particularly in the fields of (Industrial) Internet of Things, biometrics, Machine Learning and next-generation wireless systems, are opening up new opportunities for innovative authentication methods. Nevertheless, these developments come along with complexities and vulnerabilities which need to be addressed as well. Upcoming trends such as Zero Trust Architecture and Zero Knowledge Authentication are an essential step towards a stronger digital environment. These approaches are

[3] Depending on the specific application and use case; may differ

in line with the growing importance of privacy and data protection in our increasingly interconnected world. They offer robust solutions against a wide range of cyber threats, but their implementation requires a change in traditional security paradigms and user behaviour. The experience gathered demonstrates that no single authentication method can overcome all security and usability concerns: A hybrid approach combining different, situation-adapted methods to balance usability, security and privacy is certainly the most promising and effective approach. This adaptable and context-sensitive strategy is crucial in a world where digital interactions are becoming increasingly complex and integral to our daily lives. Therefore, the future of authentication depends on continuous innovation and adaptation. Building solutions that are not only technologically advanced, but also ethically sound and user-centred, will require a collaborative effort between technologists, security experts, policy makers and users. In navigating this constantly changing environment, we must continue to focus on creating a secure, efficient, and inclusive hyperconnected world for all of us.

Acknowledgement

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Förderkennzeichen 16KIS1841K ALPAKA and 16KIS1283 AI-NET PROTECT). The authors alone are responsible for the content of the paper.

References

Arp, D., Quiring, E., Pendkebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., and Rieck, K., "Lessons Learned on Machine Learning for Computer Security", *IEEE Security & Privacy, vol.* 21, no. 5, pp. 72 --77, DOI: 10.1109/MSEC.2023.3287207, 2023.

Bellovon, S. and Merritt, M., "Encrypted key exchange: password-based protocols secure against dictionary attacks", *IEEE Computer Society Symposium on Research in Security and Privacy,* Oakland, CA, USA, DOI: 10.1109/RISP.1992.213269, *1992*.

Bidgoly, A. J., Bisgoly, H. J. and Arezoumand, Z., "A survey on methods and challenges in EEG based authentication", *Computers & Security,* Band 93, DOI: 10.1016/j.cose.2020.101788, 2020.

Bonneau, J., Herley, C., van Oorschot, P. and Stajano, F., "*The Quest to Replace Passwords: A Famework for Comparative Evaluation of Web Authentication Schemes" IEEE Symposium on Security and Privacy*, San Fransisco, CA, USA, DOI: 10.1109/SP.2012.44, 2012.

Burrows, M., Abadi, M. and Needham, R., "A logic of authentication" *ACM Transactions on Computer Systems, vol.* 8, no. 1, pp. 18 -- 36, DOI: 10.1145/77648.77649, 1990.

Burr, W.E., Dodson, D.F., Newton, E.M., Perlner, R.A., Polk, W.T., Gupta, S., and Nabbus, E.A., "NIST Special Publication 800-63-2 - Electronic Authentication*", National Institute of Standards and Technology,* 2013.

Campisi, P., "Security and Privacy in Biometrics"*,* 1st Edition, Springer London, ISBN: 978-1447152293, 2013.

Chadha, A., Kumar, V., Kashyap, S. and Gupta, M., "Deepfake: An Overview", *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security,* DOI: 10.1007/978-981-16-0733-2_39, 2021.

Ciocarlie, G.F. and Zhou, J., "Securing Critical Infrastructure Across Cyber and Physical Dimensions", *IEEE Security & Privacy, vol.* 21, no. 4,, DOI: 10.1109/MSEC.2023.3282424, 2023.

Cisco, "Cisco Annual Internet Report (2018–2023)*", White Paper,* 2020.

Garman, J., "Kerberos: The Definitive Guide"*,* 1. Edition, O'Reilly & Associates, 2023.

Gassend, B., Clarke, D., van Dijk, M. and Devadas, S., "Silicon Physical Random Functions", *Proceedings of the 9th ACM Conference on Computer and Communication Security*, pp. 148 -- 160, DOI: 10.1145/586110.586132 2002.

Ho, L.K. and Katuk, N., "Social login with OAuth for mobile applications: User's view"*, IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE),* Penang, Malaysia, DOI: 10.1109/ISCAIE.2016.7575043 , 2016.

Huseynov, E. and Seigneur, J.-M., "Context-Aware Multifactor Authentication Survey", In: M. Kaufmann, Hrsg. *Computer and Information Security Handbook.* Third Edition, Boston, USA, pp. 715 -- 726, DOI: 10.1016/B978-0-12-803843-7.00050-8, 2017.

Kimovski, D., Mathá, R., Hammer, J., Mehran, N., Hellwagner, H. and Prodan, R., "Cloud, Fog, or Edge: Where to Compute?", *IEEE Internet Computing, vol.* 25, no. 4, pp. 30 -- 36, DOI: 10.1109/MIC.2021.3050613, 2021.

Lin, J., Xu, L., Liu, Y. and Zhang, X., "Composite Backdoor Attack for Deep Neural Network by Mixing Existing Benign Features", *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security,* pp. 113 -- 131, DOI: 10.1145/3372297.3423362, 2020.

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: Van Niekerk, B., Ramluckan, T. and Kushwaha, N., *Modelling Nation-state Information Warfare and Cyber-operations.* United Kingdom: Academic Conferences and Publishing International Limited, pp. 215 -- 236, ., 2022.

__, Baradie, S., Noushinfar, M., Herbst, J., Weinand, A., and Schotten, H.D., "Towards the Sixth Generation (6G) Wireless Systems: Thoughts on Physical Layer Security", *Mobile Communication - Technologies and Applications - 25. VDE/ITG Fachtagung Mobilkommunikation*, Osnabrück, Germany, 2021.

__, Herbst, J., Klingel, S., Franke, S., Wolff, A., Rüb, M., Reddy, R., Rahm, M., and Schotten, H.D., "Connectivity in the era of the (I)IoT: About Security, Features and Limiting Factors of Reconfigurable Intelligent Surfaces", *Discover the Internet of Things, vol.* 3, no. 16, DOI: 10.1007/s43926-023-00046-1, 2023.

__, 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, TN, USA, DOI: 10.34190/IWS.21.016, 2021.

__, and Schotten, H.D., "Physical Layer Security: About Humans, Machines and the Transmission Channel", *21st European Conference on Cyber Warfare and Security (ECCWS2022),* Chester, United Kingdom, DOI: 10.34190/eccws.21.1.403, 2022.

__, Tjabben, A., Rüb, M., Herbst, J., Sanon, S.P., Reddy, R., Munoz, Y., Schotten, H.D., "Designing Security for the Sixth Generation: About Necessity, Concepts and Opportunities", *22nd European Conference on Cyber Warfare and Security*, Academic Conferences International, Athens, Greece, DOI: 10.34190/eccws.22.1.1207, 2023.__, 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, DOI: 10.1109/ICDIS.2018.00013, 2018.

Ma, C., Li, J., Wei, K., Liu, B., Ding, M., Yuan, L., Han, Z., and Poor, H.V., "Trusted AI in Multiagent Systems: An Overview of Privacy and Security for Distributed Learning", *Proceedings of the IEEE, vol.* 111, no. 9, pp. 1097 -- 1132, DOI: 10.1109/JPROC.2023.3306773, 2023.

Needham, R.M. and Schroeder, M.D., "Using encryption for authentication in large networks of computers", *Communications of the ACM, vol.* 21, no. 12, pp. 993 -- 999, DOI: 10.1145/359657.359659, 1978.

O'Gorman, L., "Comparing passwords, tokens, and biometrics for user authentication" *Proceedings of the IEEE, vol.* 91, no. 12, pp. 2021 -- 2040, DOI: 10.1109/JPROC.2003.819611, 2003.

Otway, D. and Rees, O., "Efficient and timely mutual authentication", *ACM SIGOPS Operating Systems Review, vol.* 21, no. 1, pp. 8 -- 10, DOI: 10.1145/24592.24594, 1987.

Radha, V. and Reddy, D.H., "A Survey on Single Sign-On Techniques", *Procedia Technology,* Band 4, pp. 134 -- 139, DOI: 10.1016/j.protcy.2012.05.019, 2012.

Ruzomberka, E., Love, D.J., Brinton, C.G., Gupta, A., Wang, C.-C., and Poor, H.V., "Challenges and Opportunities for Beyond-5G Wireless Security", *IEEE Security & Privacy, vol.* 21, no. 5, pp. 55-66, DOI: 10.1109/MSEC.2023.3251888 2023.

Sanon, S. P., Lipps, C. and Schotten, H.D., "Fully Homomorphic Encryption: Precision Loss in Wireless Mobile Communication", *Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit),* Gothenburg, Sweden, DOI: 10.1109/EuCNC/6GSummit58263.2023.10188286, 2023.

Sarkar, A., Abbott, A.L. and Doerzaph, Z., "*Biometric authentication using photoplethysmography signals", IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS),* Niagrara Falls, NY, USA, DOI: 10.1109/BTAS.2016.7791193, 2016.

Schneier, B., "Secrets & Lies - Digital Security in a Networked World",. 1st Edition, New York: John Wiley & Sons, Inc., 2000.

__, "Liars & Outliers - Enabling the Trust that Society need to thrieve", 1st Edition, Indianapols: John Wiley & Sons, Inc., 2012.

__, "Carry On: Sound Advice from Schneier on Security". 1st Edition, Wiley, 2014.

__, "Applied Cryptography: Protocols, Algorithms and Source Code in C", 25th Anniversary Edition, John Wiley & Sons Inc., 2015.

Smart, N., "Computing on Encrypted Data", *IEEE Security & Privacy, vol.* 21, no. 4, pp. 94-98, DOI: 10.1109/MSEC.2023.3279517, 2023.

Tolanur, J. and Chaudhari, S., "DDoS Attacks Analysis with Cyber Data Forensics using Weighted Logistic Regression and Random Forest", *International Conference on Device Intelligence, Computing and Communication Technologies (DICCT),* Dehradun, India, DOI: 10.1109/DICCT56244.2023.10110133 , 2023.

Tolpegin, V., Truex, S., Gursoy, M.E. and Liu, L., "Data Poisoning Attacks Against Federated Learning Systems", *European Symposium on Research in Computer Security,* pp. 480 -- 501, 2020.

Tuptuk, N., Hazell, P., Watson, J. and Hailes, S., "A Systematic Review of the State of Cyber-Security in Water Systems", *Water, vol.* 13, no. 81, DOI: 10.3390/w13010081, 2021.

Wei, Z., Wang, Z., Zhang, J., Li., Q., Zhang, J., and Fu., H.Y., "Evolution of optical wireless communication for B5G/6G", *Progress in Quantum Electronics,* Band 83, DOI: 10.1016/j.pquantelec.2022.100398, 2022.

Human-Centred AI in Military Cyber Operations

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

clara.maathuis@ou.nl

Abstract: Military Cyber Operations are an integral part of modern warfare and national security strategies as they crossed the science fiction realm, represent a real operational battlefield, and developed into an option in military toolboxes. Seeing ongoing technological advancements that allow the creation and use of complex mechanisms and technologies, the increasing digitalization of critical infrastructure, the growing abundance of data collected, generated, and exchanged between multiple parties, and the rise of stakeholders engaging in building and/or executing military Cyber Operations together with the increased number of such operations being conducted all over the globe reflects important lessons that need to be learned: the core element of such operations are humans: they build, acquire, execute, and assess them while also being impacted entities by them, e.g., through (psychological or physical) injury or death, or damage or destructions of human infrastructure. Moreover, the key process governing all the life cycle phases of military Cyber Operations is human decision-making or humanly assisted/augmented decision-making relying on advanced intelligent methods built with AI. Nevertheless, building and conducting military Cyber Operations should be done in a legal, responsible, and effective way implying a deep understanding of the context and adversary, proper target and cyber weapon selection, development, and use, and a clear overview of potential effects produced. These represent important aspects that should be properly defined and tackled in this domain. Hence, this research aims to introduce the Human-Centred AI concept and approach in the military cyber domain to illustrate ways to prioritize human involvement and interaction, human understanding, effective decision-making, and ethical considerations when building and conducting military Cyber Operations. To this end, an extensive literature review is conducted in the military, cyber, and AI domains together with instantiation on military Cyber Operations.

Keywords: Human-centred AI, Trustworthy AI, Responsible AI, Cyber operations, Military operations

1. Introduction

motto: "Machines take me by surprise with great frequency." (Alan Turing)

AI emerged as a transformative force in the military domain, redefining the nature of modern warfare. Its unmatched capacity to process massive datasets and combine it with field knowledge and expertise empowers situational awareness, facilitates data, knowledge, or hybrid-driven decision-making, and optimizes resource allocation, not only in traditional military operations but also in the intricate military cyber domain. In the cyberspace battlefield, military Cyber Operations are conducted as independent or part of broader military operations to achieve military goals against adversaries using cyber weapons/capabilities (Maathuis, Pieters & van den Berg, 2018a). Nevertheless, to integrate AI in this domain and benefit from its advantages while dealing with its challenges, the stakeholders involved need to assure that AI not only that it supports or enhances human decision-making (Ahmed, 2022), but that is able to facilitate cyber capabilities that are safe, responsible, and trustworthy, thus human centered. This can be done by assuring a harmonious balance between harnessing AI's potential and upholding international social and ethical norms and values since the design of AI systems and going forward in their life cycle when building and conducting military Cyber Operations while actively involving the stakeholders in this process through a socio-technical approach (Makarius et al., 2020). This perspective is called Human Centered AI (HCAI), and in the AI domain as well as direct critical application domains like defense, this area is surrounded by open questions and challenges (Vorm, 2020). Therein, humans are placed at the core of decision-making processes while AI and human agents form teams for achieving well-defined objectives and adhering to socio-ethical norms and values like international law, transparency, accountability, and safety (Shneiderman, 2020a).

Nonetheless, the focus on employing a HCAI approach within military Cyber Operations highlights a critical knowledge gap that requires immediate consideration from both researchers and practitioner efforts. This calls for creating and building corresponding concepts, methodologies, and techniques, effectively bridging the gap between theory and practical implementation. Addressing this knowledge gap is essential to unlock the full potential of AI techniques while upholding social and ethical norms, principles, and values. In pursuit of this objective, this research seeks to establish a clear understanding of the HCAI concept within this context and introduce a design framework for its responsible development and use. To this end, a transdisciplinary research approach is employed, drawing upon insights from the fields of AI, AI ethics, military studies, cybersecurity, and military Cyber Operations. This approach follows the Design Science Research methodology while incorporating perspectives rooted in the Values Sensitive Design approach. Accordingly, a

comprehensive literature review is conducted, complemented by illustrative scenarios for exemplifying the assessment of proportionality in military Cyber Operations, serving as concrete instantiation to support both the definition and framework proposed.

The outline of this article is organized as follows. Section 2 sets the background of this research and discusses related studies. Section 3 presents the definition proposed for HCAI in military Cyber Operations. Section 4 advances a design framework for HCAI in this domain with exemplification on the proportionality principle. At the end, Section 5 briefly discusses the findings and future perspectives of this research.

2. Research Approach and Background

This research aims to define the concept of HCAI in the military Cyber Operations domain and propose a framework for building and using HCAI systems in a responsible way. Accordingly, the following research questions are formulated:

- How to define the HCAI concept in the context of military Cyber Operations?
- How to design a framework for building and using HCAI systems in military Cyber Operations in a responsible way?

To achieve this aim, a transdisciplinary research approach is considered by including knowledge, methods, and techniques from the AI, AI ethics, military, cyber security, and military Cyber Operations domains using the Design Science Research methodology while respecting the Values Sensitive Design perspective (Peffers et al., 2007; Umbrello & Van de Poel, 2021; Maathuis, 2022a). An extensive literature review is conducted in the ACM Digital Library, IEEE Digital Library, Scopus, and Google Scholar scientific databases using combinations of keywords like military operations, Cyber Operations, human centered, and HCAI.

A socio-technical approach is promoted by (Makarius et al., 2020) where the authors state that bringing AI into organizations, systems, and processes implies understanding the novelty of AI systems and its scope dimensions. The authors argue that AI positions itself apart from its technological predecessors as it implies interaction between the physical, digital, and biological worlds while implying cognitive, relational, and structural issues like strategic decision making and knowledge sharing, trust and identification, coordination, training and development, and socialization. Sperrle et al. (2021) provide a comprehensive overview of evaluations for human-centred machine learning focusing on human-related factors that influence trust, interpretability, and explainability. Usmani, Happonen & Watada (2023) argue that HCAI systems should be developed for working together with humans, involving them in the decision-making processes, and leverage human expertise, intuition, and creativity through ethical considerations and effective human-AI collaboration.

Van den Bosch (2018) identify three characteristics of successful joint activities in human-AI collaboration: be mutually predictable in their actions, be mutually directable, and maintain common ground. Geng & Varshney (2019) model human-machine collaborative decision-making in settings that include random local thresholds, decision fusion in integrated human-machine networks, and binary decision making under cognitive biases. (Shneiderman, 2020a) propose a generic framework for HCAI for building reliable, safe, and trustworthy systems upon the following pillars adopted in the present research: design for high level of human control and high levels of computer automation for increasing human performance, understand the situations where full human control or full computer control and necessary, and avoid the dangers of excessive human control or excessive computer control. The classical OODA (Observe, Orient, Decide, Act) loop is used by integrating the following elements: human values and knowledge, trust and needs, capabilities and intentions, and environment, AI internal model, and the external world. In the same lines, Shneiderman (2020b) provides a series of guidelines for building reliable, safe, and trustworthy HCAI systems, i.e., sure that the systems are reliable based on sound software engineering practices, promote a safety culture through business management strategies, and trustworthy certification assured through independent oversight. Cronholm & Göbel (2020) gather design principles for developing HCAI solutions in decision-making processes: (i) design for amplified decision-making and possible responses from both humans and AI, i.e., a positive response from human and AI, a negative response from both human and AI, a positive response from human and a negative response from AI, and a negative response from human and a positive response from AI, (ii) design for unbiased decision-making through the involvement of several roles and consensus-based decision-making by including multiple perspectives for reducing biases, uncertainty, and vagueness of human judgement, and (iii) design for human and AI learning through procedural support that enables both human and machine learning which will improve the overall performance and capability of the system.

Ozmen Garibay et al. (2023) identify the following six challenges for building HCAI systems: human well-being that asks for considering the unique features of AI that might impact well-being while applying foundational and actionable characteristics of well-being oriented AI; responsible design of AI including values like transparency and fairness; privacy arguing for the right to be alone; limit access to the self/shields self from access, secrecy, and personhood; design and evaluation framework; governance and independent oversight; and human-AI interaction. For autonomous driving, De Caro et al. (2022) proposes an AI-as-a-Service toolkit for human-centred intelligence integrating a data gathering ecosystem from wearables for automatizing stress recognition. The core element is human behaviour, which is modelled in relation to vehicle's state, environment's state, and driver's state.

For building trustworthy robots and autonomous systems, He et al. (2021) investigate challenges for HCAI like safety when dealing with dynamic environments, and security issues from unknown software, hardware, and communication vulnerabilities. The authors call for tackling these issues by adopting a temporal vision that focuses on immediate, near-and-medium-term, and longer-term concerns. In the military domain, (He et al., 2023) identify the following open challenges for implementing HCAI solutions: the use of automated planning algorithms to aid Commanders in complex and time-critical decisions, the need for efforts that enable huma users to provide feedback in ways that are natural, flexible, and that result in proper updates and future learning for the algorithms, and the use of predictive dashboards for unit readiness in activities like acquisition program planning, manning, and recruiting. Ahmed (2022) investigate the role of ML models and the collaboration between humans and ML models during activities taken in the military intelligence operation process, e.g., when establishing intelligence requirements, during estimation and analysis, and preparing intelligence reports. Such activities are integrated when building and deploying, for instance, video surveillance and acoustic detector systems.

In the cyber security domain, Ansari et al. (2018) build a problem-solving ontology for cyber physical production systems identifying as problem characteristics: isolation (unconnected), connected (related), repetitive, unique (singular), transformable into a standard problem, ex ante (based on forecasts), and ex post (based on actual results). The study identifies complementary situations between humans and machines like fulfilment of various tasks, information processing, problem solving, and decision making. Dustdar, Nastic & Scekic (2016) bring a new vision of cyber-human smart city built upon a values architecture structured as follows: core technologies, enablers (provisioning and governance, incentive management, activity coordination and social orchestration, and monitoring with data analytics), value generation, added-value services, and smart city goals (quality of life, sustainability, and development). Maathuis (2023) introduces the concept of Human Centred Explainable AI (HCXAI) in the military cyber domain and proposes a framework for building and integrating HCXAI solutions in military Cyber Operations.

Governance efforts are also directed to HCAI by different stakeholders. The EU Commission published its vision on building trust in human-centric AI stressing the consideration of the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights. Accordingly, the following seven requirements should be accounted for developing and achieving trustworthy AI: 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 (EU Commission, 2019). Moreover, (EU Commission, 2023) provides a clear overview of efforts for building HCAI solutions by EU countries while stressing the importance of respecting and implementing social and ethical norms and values like transparency, safety, privacy, and inclusion. UNESCO argues that for ensuring ethical AI development and use, laws, cooperation and collaboration, impact assessment, governance mechanisms should be considered (UNESCO, 2021). IEEE, through its Global AI systems (AIS) well-being initiative (IEEE, 2020) aims to propose recommendations and resources for building AI systems now and in the future that align and prioritize the human well-being and ecological sustainability.

This extensive review illustrates that several research, practitioner, and governmental efforts recently started for building and using HCAI systems in a responsible way. Nevertheless, a limited range of studies were conducted in the military and cyber domains, while in the domain of military Cyber Operations this represents a knowledge gap that this research aims to tackle.

3. Definition

HCAI represents a paradigm shift in the AI realm and recently became an important AI research area as it aims at building system that serve a greater purpose than just innovation – they are directed to supporting or

enhancing the overall human experience. This implies that when AI systems encounter humans, they need to understand how humans behave and what they want while humans need to make sure that AI systems embed methods that help humans understanding them (Riedl, 2019). Humans and AI become teammates where AI provides input to humans to make decisions and act (bring AI in the loop of human intelligence) while human intelligence is fundamental to building AI (human intelligence in the AI loop): this relates to combining human intelligence with AI into hybrid intelligence for reaching superior results then could have accomplished separately (Dellermann et al., 2019). Moreover, HCAI implies a socio-technical perspective and defines systems that continuously improve (Cronholm & Göbel, 2019) focusing not only on human input, but also on interactions, collaboration, and the involvement of diverse stakeholders while enabling feedback and model development (Sperrle et al., 2021). Nowak, Lukowicz & Horodecki (2019) relate HCAI to reasoning implying that HCAI needs to understand human reasoning through morals, motivations, and emotions.

The stakeholders involved consider a user-centered participatory design methods for engagement with diverse stakeholders while putting humans at the center of design, thinking, emphasizing user experience design by measuring aspects like satisfaction, interests, needs, and ensuring meaningful control (Shneiderman, 2020b; Mhlanga, 2022). This would facilitate benefits like making well-informed decisions, scalability, and dependability, building highly effective software and products, increasing inclusivity, and customized user experiences. Schmager & Vassilakopoulou (2023 propose the following HCAI guidelines: build reliable and transparent systems based on software engineering practices, pursue safety through effective business management strategies, and increase trust through certification and independent oversight to assure transparency and accountability enhance innovation, public confidence, and societal values.

Table 1: Key elements in HCAI definitions

Key element in existing HCAI definitions	Source
Contact	Riedl (2019)
Connection	Dellermann et al. (2019)
Engagement	Shneiderman (2020b)
Reliability, transparency, safety, and trust	Schmager & Vassilakopoulou (2023)
Decision, control, inclusion	Mhlanga (2022)
Socio-technical perspective	Cronholm & Göbel (2022)
Feedback	Sperrle et al., (2021)
Reasoning	Nowak, Lukowicz & Horodecki (2019)

These definitions tackle important features or key elements that characterize HCAI and are depicted in Table 1. As they provide a general perspective of this concept, for making sure that the development of HCAI systems is done in a responsible way while considering their possible impact and for preventing confusion between the stakeholders involved, this concept is tailored and defined in the military domain as follows:

HCAI in MCO = a sub-field of AI that deals with centring human needs, reasoning, input, control, and feedback 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.

Its elements are defined as follows:

- *Centring human needs, reasoning, input, control, and feedback* capture the major socio-technological features that characterize HCAI.
- *Design, development, deployment, and use* of AI systems refer to the life cycle phases of AI systems considering a software engineering approach.
- *Military Cyber Operations* are independent military operations or part of other broader military operations that make use of cyber weapons/capabilities for reaching military goals (Maathuis, Pieters & Van den Berg, 2018b).

4. Framework

The proposed definition brings together the socio-technical elements of this concept and paradigm. This holistic approach positions humans in a dimension where AI is a trusted ally helping human stakeholders to navigate and tackle the complexities of this domain while achieving their goals and preserving the socio-ethical norms and values. For enhancing awareness, trust, and resilience of stakeholders and AI systems involved when conducting military Cyber Operations, the HCAI triad and framework are further introduced and

presented in Figure 1, containing three layers (technical, socio-technical, and social) guided by three core actions of building intelligent systems (understanding, building, and deciding) while embedding human-interaction ke- elements (involvement, interaction, and control).

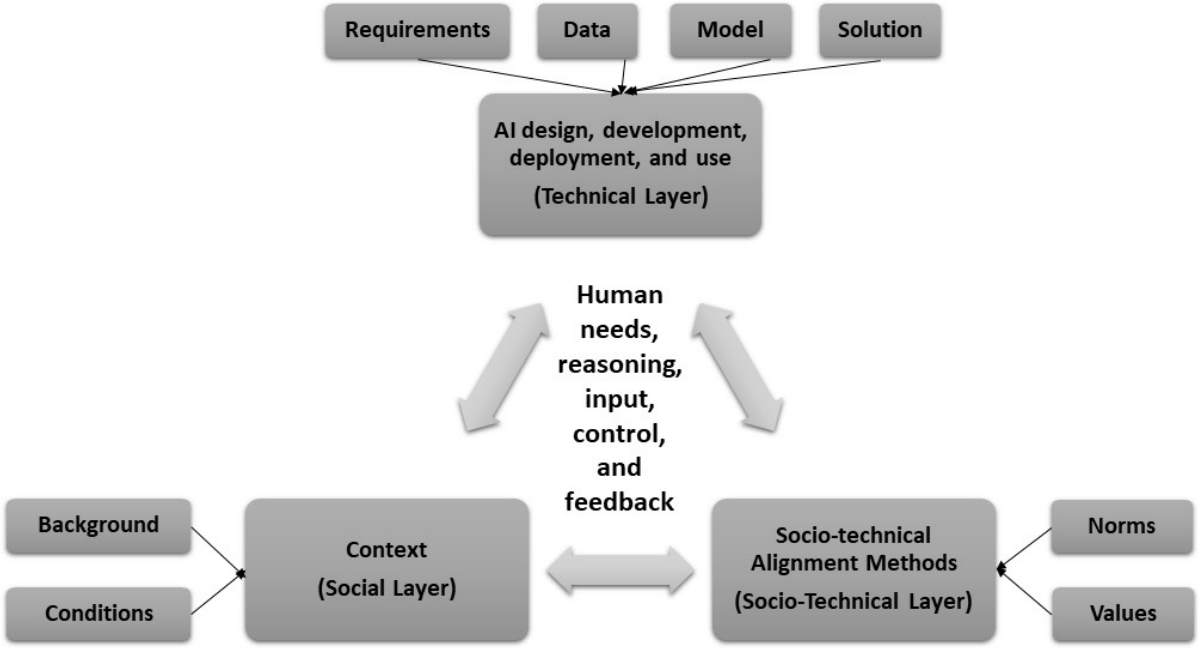


Figure1: The HCAI triad and framework

Human needs, reasoning, input, control, and feedback core which represent the core of HCAI being expressed by (Ansari et al., 2018; Shneiderman, 2020a; Maathuis, 2023; Manzoor et al., 2023):

- *needs* that include human aims, vision, experiences, and expectations of stakeholders involved. These should be directly implemented in context conditions and technical requirements, and be aligned with socio-ethical norms and values.
- *reasoning* that enables AI systems to navigate, capture, and model human needs, moral dilemmas, and make choices that prioritize human well-being and achievement of the objectives defined.
- *input* which can be provided in all phases of AI systems development, thus from defining the objective and design up to the deployment of AI systems. Given current and future developments in the field of generative AI, the input is multimodal, i.e., has different types of data (e.g., text, image, audio) that the model can understand, process, or further generate.
- *control* which embeds the level of control that humans have when building and using AI systems thinking in a scheme of problem and solution profile, problem-solver profile, and problem-solving profile that includes different levels of autonomy, from no autonomy to partial or conditional, to full autonomy while including different forms and characteristics of assistance, e.g., warnings, indicators, and notifications. In an ideal situation, building a reliable, safe, and trustworthy AI system implies having a high level of human control while having a high level of automation.
- *feedback* implies central role as it sustains continuous improvement, the effectiveness of AI systems, and assures the alignment between AI and human needs plus socio-ethical norms and values.

A direct example that can be considered is the proportionality assessment in military Cyber Operations (Maathuis & Chockalingam, 2023). The stated military objectives to be achieved *(i.e., needs)* are merged with assessments for components like collateral damage to assure that the unintended effects of using a specific cyber weapon (i.e., malware) would not be excessive in relation with the anticipated military advantage *(i.e., reasoning).* Accordingly, different types and sources of data are used (i.e., input) for aspects like target geolocation and cyber weapon characteristics. Concrete levels of autonomy and interaction of the AI-based cyber weapon are established since its design phase and corresponding methods and techniques are implemented to build a reliable, safe, and trustworthy AI solution while assuring real interaction between it and the stakeholders involved (i.e., control and feedback) for aspects like vulnerability identification, weapon extraction, and impact assessment.

Context which represents the social layer where the environment, and circumstances of building the AI system are considered (Makarius et al., 2020; He et al., 2021; Gao et al., 2021):

- *background* includes information about the military Cyber Operation, current strategic, economic, civilian, cultural, and media situations, previous lessons learned from former assessments, strategic, operational, and tactical requirements, and information about the stakeholders involved or possibly impacted through its execution.
- *condition*s refer to specific aspects or requirements that should be accounted when building the AI systems and conducting the military Cyber Operation, e.g., ROE and ISR reports, to assure that informed decisions are made in an adaptive manner taking into consideration the dynamism and uncertainty surrounding this battlefield.

For the same example, previous insights about civilian objects in the area and current information about a zero-day vulnerability exist *(i.e., background)* and can be further used while respecting the precise military-legal and political conditions stated in the corresponding ROE defined for this operation *(i.e., conditions).*

Socio-technical alignment methods which represents the socio-technical layer where social-ethical norms and values are included to embed the implementation of ethical standards, legal compliance, and codes of conduct of AI systems used (Ansari et al., 2018; Shneiderman, 2020a; Shneiderman, 2020b; Dustdar, Nastic & Scekic, 2016; Chockalingam & Maathuis, 2022):

- *norms* refer to the adherence to applicable laws and regulations that govern the development and use of AI systems and data privacy plus military cyber professional codes of conduct established by the stakeholders involved.
- *values* point to respecting ethical and moral aspects like transparency, accountability, fairness, beneficence and non-maleficence, privacy, and fairness.

In the example above, IHL compliance is mandatory by respecting the principles of distinction, precaution, and proportionality while respecting Human Rights by protecting civilians and the right of life *(i.e., norms).* Accordingly, dedicated efforts should be developed for engaging the desired military target with the cyber weapon that would not produce disproportionality and would make sure that the right of life is guaranteed while the avoidance or minimization of collateral damage is assured before its use *(i.e., values).*

AI design, development, and use which represents the technical layer that encompasses the AI life cycle phases used in a military Cyber Operation including the following core elements (Maathuis, 2022b):

- *requirements* are technical functional and non-functional guidelines for stakeholders involved and assure the successful development and deployment of AI systems.
- *data* points to multi-source and multi-nature data, real from the field or previous incidents, simulated from current mirroring settings or previous operations, or synthetically generated based on well-established profiles.
- *model* represents the AI model that can imply either a knowledge-based, data-based, or hybrid AI-based approach by relying just on knowledge or expertise, data, or a combination thereof.
- *solution* refers to the specific AI system that will be integrated and used in a military Cyber Operation.

For the same example, the technical requirements are established in terms of infrastructure at all levels, i.e., software, hardware, communication lines, and data (i.e., requirements and data), the paradigms, methods, and techniques are selected for designing the AI system (i.e., model) which is further built and prepared for use for assessing the proportionality in military Cyber Operations while assuring the avoidance or reduction of collateral damage (i.e., solution).

5. Conclusions

By prioritizing the development and integration of AI systems in the military domain that are designed with a deep understanding and centralization of human aims, needs, capabilities, and ethical considerations, the stakeholders involved when building and using them can ensure that these systems become force multipliers rather than risks (Vorm, 2020). Taking into consideration that the elements of the cyberspace battlefield are in a continuous movement and the adversary's capabilities constantly evolve, plus the agility, precision, and effectiveness offered by AI systems when building and conducting military Cyber Operations are invaluable (Usmani, Happonen & Watada, 2023; Maathuis, Pieters & van den Berg, 2018c). For instance, enhanced cyber threat detection, improved overall resilience to unknown cyber and even hybrid threats, and optimization of

cyber weapons effectiveness while minimizing its unintended effects while fostering values like transparency, trust, and safety all along with centering on the stakeholders involved and/or impacted by their actions, facilitates, or strengthens the successful collaboration between humans and AI systems on this battlefield as fundament of assuring a reliable, safe, responsible, and trustworthy human-machine partnership. Nevertheless, the emphasis on a HCAI approach in military Cyber Operations underscores a crucial knowledge gap that demands urgent attention from researchers and practitioners through a bridge between theory and practice by developing corresponding artefacts-concepts, methods, and techniques.

Tackling this knowledge gap is imperative to harness the full potential of AI while safeguarding the principles of humanity and socio-ethical norms, principles, and values in military Cyber Operations. Accordingly, this research aims to define the concept of HCAI in this context and propose a corresponding responsible framework for its development and use. To achieve these goals, a transdisciplinary research approach combines knowledge and methods from the AI, AI ethics, military, cyber security, and military Cyber Operations domains, using the Design Science Research methodology while incorporating the Values Sensitive Design perspective. Extensive literature review is conducted in combination with illustrative scenarios on the proportionality assessment in military Cyber Operations provided as exemplification for the definition and framework proposed. This research continues by integrating other aspects like cognitive and psychological ones in the decision-making processes, and by analyzing mechanisms for facilitating cross-domain collaboration and human-AI teaming for building responsible, safe, trustworthy military Cyber Operations.

References

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.

Ansari, F., Khobreh, M., Seidenberg, U., & Sihn, W. (2018). A problem-solving ontology for human-centered cyber physical production systems. *CIRP Journal of Manufacturing Science and Technology*, *22*, 91-106.

Chockalingam, S., & Maathuis, C. (2022). An ontology for effective security incident management. In *International Conference on Cyber Warfare and Security* (Vol. 17, No. 1, pp. 26-35). Academic Conferences International Limited.

Cronholm, S., & Göbel, H. (2022). Design principles for human-centred AI.

De Caro, V., Bano, S., Machumilane, A., Gotta, A., Cassarà, P., Carta, A., ... & Bacciu, D. (2022). AI-as-a-Service Toolkit for Human-Centered Intelligence in Autonomous Driving. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)* (pp. 91-93). IEEE.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*, 637-643.

Dustdar, S., Nastic, S., & Scekic, O. (2016). A novel vision of cyber-human smart city. In *2016 fourth ieee workshop on hot topics in web systems and technologies (hotweb)* (pp. 42-47). IEEE.

EU Commission (2019). Building Trust in Human-Centric Artificial Intelligence.

E.U. Commission (2023). EU AI Act: The European approach to AI.

Geng, B., & Varshney, P. K. (2019). On decision making in human-machine networks. In *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)* (pp. 37-45). IEEE.

He, H., Gray, J., Cangelosi, A., Meng, Q., McGinnity, T. M., & Mehnen, J. (2021). The challenges and opportunities of human-centered AI for trustworthy robots and autonomous systems. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(4), 1398-1412.

IEEE (2020). Global AI systems (AIS) well-being initiative.

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). Assessment methodology for collateral damage and military (Dis) Advantage in cyber operations. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)* (pp. 1-6). IEEE.

Maathuis, C., Pieters, W., & van den Berg, J. (2018c). 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).

Maathuis, C. (2022a). 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. (2022b). 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. (2023). Human Centered Explainable AI Framework for Military Cyber Operations. In *MILCOM 2023 IEEE Military Communications Conference (MILCOM)*. IEEE.

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262-273.

Manzoor, M. A., Albarri, S., Xian, Z., Meng, Z., Nakov, P., & Liang, S. (2023). Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications. *arXiv preprint arXiv:2302.00389.*

Mhlanga, D. (2022). Human-centered artificial intelligence: the superlative approach to achieve sustainable development goals in the fourth industrial revolution. *Sustainability*, *14*(13), 7804.

Nowak, A., Lukowicz, P., & Horodecki, P. (2018). Assessing artificial intelligence for humanity: Will ai be the our biggest ever advance? or the biggest threat [opinion]. *IEEE Technology and Society Magazine*, *37*(4), 26-34.

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., ... & Xu, W. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, *39*(3), 391-437.

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.

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human behavior and emerging technologies*, *1*(1), 33-36.

Schmager, S., P., I., & Vassilakopoulou, P. (2023). Defining Human-Centered AI: a Comprehensive Review of HCAI Literature.

Shneiderman, B. (2020a). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504.

Shneiderman, B. (2020b). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1-31.

Sperrle, F., El-Assady, M., Guo, G., Borgo, R., Chau, D. H., Endert, A., & Keim, D. (2021). A Survey of Human-Centered Evaluations in Human-Centered Machine Learning. In *Computer Graphics Forum* (Vol. 40, No. 3, pp. 543-568).

Umbrello, S., & Van de Poel, I. (2021). Mapping value sensitive design onto AI for social good principles. *AI and Ethics*, *1*(3), 283-296.

Usmani, U. A., Happonen, A., & Watada, J. (2023). Human-Centered Artificial Intelligence: Designing for User Empowerment and Ethical Considerations. In *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)* (pp. 01-05). IEEE.

UNESCO (2021). The Unesco recommendations on the ethics of AI: shaping the future of AI: shaping the future of our societies.

Van Den Bosch, K., & Bronkhorst, A. (2018, May). Human-AI cooperation to benefit military decision making. NATO.

Vorm, E. S. (2020). Computer-centered humans: why human-AI interaction research will be critical to successful AI integration in the DoD. *IEEE Intelligent Systems*, *35*(4), 112-116.

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.

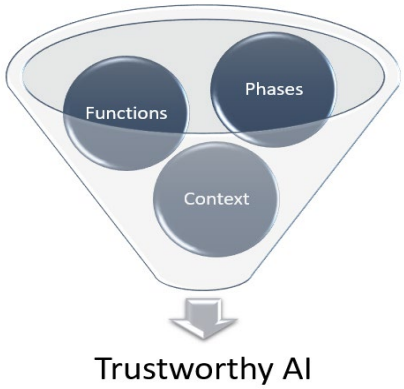


Figure 1: Dimensions of Trustworthy AI in military Cyber Operations

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.

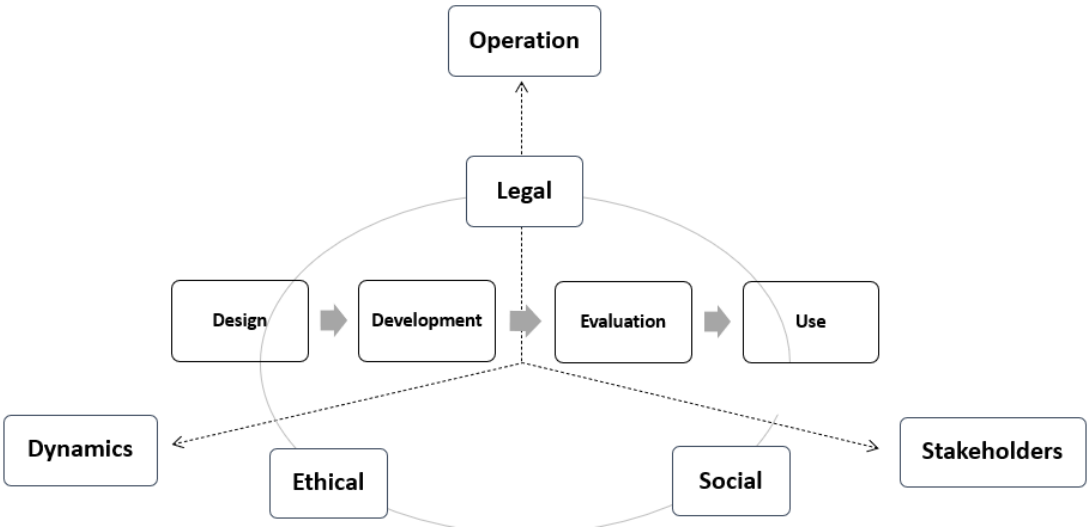


Figure 2: Trustworthy AI Framework in military Cyber Operations

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.

Ukrainian Thoughts and Feelings Based on One Year War Content Analysis

Clara Maathuis[1] and Iddo Kerkhof[2]
[1]Open University of the Netherlands, Heerlen, The Netherlands
[2]Independent Researcher, The Netherlands

clara.maathuis@ou.nl
email@iddo.eu

Abstract: In the age of remarkable technological innovations such as AI, quantum computing, IoT, and blockchain, one could argue that the world is becoming an increasingly connected, advanced, enlightened, and peaceful place. Nevertheless, paradoxically, conflicts and wars are still born deeply rooted in complex strategic, historical, and economic dynamics, and continue to arise finding a direct representation in both physical and digital battlefields through powerful and dynamic environments like social media. This environment enabled unprecedented connectivity and ideas, thoughts, ideologies, and experiences exchange shaping the narrative, amplifying ideas, feelings, and experiences while directly addressing the underlying causes, dynamics, and implications of conflicts and war through a multifaceted and multi-nuanced approach. The outbreak and ongoing war in Ukraine took the international community off guard representing a significant turning point in the contemporary era that received immediately intensive global attention, united global humanitarian, and strategic efforts to support and help the ones in need, and reflected its geopolitical and economic global complexities, dynamics, and implications. At the same time, a high degree of disinformation and misinformation exists surrounding different aspects about this war e.g., actions taken, and impact produced on civilians, which contributes to creating an altered perspective of reality. While different conventional media and social media outlets together with research, governmental, and practitioner efforts revealed and reflected on the nature, major incidents, and their known impact on civilians and civilian objects as well as military personnel and military objects, yet limited attention and a lower voice is given through dedicated studies to thoughts and emotions of Ukrainian users in unconventional social media platforms like Telegram. This represents the knowledge gap tackled in this research by building a set of Machine Learning-based solutions for analysing the discourses and sentiments of Ukrainian users using the Design Science Research methodology in a Data Science research approach. From this analysis, reflection is provided to important days and incidents experienced by Ukrainian population since February 2022 in the first year of war.

Keywords: Ukraine war, Russia-Ukraine war, Topic modelling, Sentiment analysis, Machine learning

1. Introduction

> *"Before a war military science seems a real science, like astronomy; but after a war it seems more like astrology." (Rebecca West)*

The ongoing war in Ukraine continues to send shockwaves through the global conscience and profound political, ideological, social, and economic consequences are reverberated far beyond the battlefield, forcing nations to rethink their alliances, strategies, policies, and programs. In this digital age, this war represents a tumultuous transformation where social media platforms together with their discourses, experiences, and emotions serve as both catalyst and chronicler. Yet, through viral posts and real-time content, social media manipulation mechanisms like disinformation and misinformation are intertwined in a tumultuous dance by spreading next to real also altered war insights, perspectives, struggles, and opinions (Fard & Maathuis, 2021; Agarwal, Punn & Sonbhadra, 2022; Sedrakyan, 2022). Hence, the urgency of safeguarding cyberspace by illustrating real aspects characterizing discourses, experiences, and feelings of Ukrainian people directly involved in this conflict are critical for discerning truth from falsehood and for understanding the impact of these battles beyond the physical battlefield (Alibudbud, 2022; Maschmeyer et al., 2023). During this war, Telegram represents the unconventional social media platform that most quickly rose in popularity and use due to its various channels (Nazaruk, 2022) and its power to provide the latest information about various war aspects (Ptaszek, Yuskiv & Khomych, 2023). However, while dedicated efforts to understanding discourses and emotions of users in other social media platforms like Twitter exist, a limited body of studies is directed to unconventional social media platforms like Telegram and TikTok (Zhu et al., 2022). This creates a gap in understanding these through such platforms (Steel, Parker & Ruths, 2023) and without tackling them, the spread of social media manipulation increases (Ye et al., 2023).

Hence, the aim of this research is to capture essential topics tackled in Ukrainian discourses together with corresponding sentiments in Telegram in the first year of war, between February 24, 2022, to February 23, 2023, and through this approach to contribute to reducing or countering ongoing social media manipulation campaigns that focus on spreading disinformation and misinformation about this war. Accordingly, the main

research question formulated is: What are the main discourses and corresponding sentiments of Ukrainian Telegram users in the first year of war? Building on previous efforts in a transdisciplinary context, the Design Science Research Methodology is followed in a Data Science approach (Shearer, 2000; Peffers et al., 2007). Accordingly, the contributions of this research are defined as follows:

- A series of Machine Learning models for extraction and analysis of topics and sentiments based on a Telegram dataset with Ukrainian messages from the first year of war.
- A basis for enhancing social media manipulation countering efforts against disinformation and misinformation on platforms that receive less research attention like Telegram, representing a call for further joint research and practitioner efforts on this behalf.

The remainder of this article is structured as follows. Section 2 discusses the background and important relevant studies regarding related datasets and methods built for understanding diverse dimensions that characterise the discourses and dynamics of the ongoing war in Ukraine. Section 3 presents the methodological approach as well as the implementation choices taken in this research. Section 4 shows the results obtained tackling both topics discussed and corresponding sentiments of users. Conclusively, Section 5 provides concluding remarks and future research perspectives.

2. Related Research

Sedrakyan (2022) analyse the implications of sanctions applied to Russia between 2014-2018 as a response to political instability in Ukraine, annexation of Crimea, human rights violation in Russia, and interference in the US 2016 presidential elections. The dataset contains sanctions applied by Western and US countries in relation to their impact on transition economies. Alibudbud (2022) investigates trauma-related aspects regarding PTSD (Post-Traumatic Stress Disorder) of Ukrainian people based on Google trends. While the authors are focusing on data between January to March 2022 and include main conflict areas at that time, the results show an increase by almost 50% in PTST-related searches and show that humanitarian efforts need to include PTSD-related services and that substantial public concern should also be directed to PTSD screening and services. At the same time, Khalfaoui, Gozgor & Goodell (2023) stress the dependency between cryptocurrencies and investment horizon and market state, and further examine the impact (negative or positive) of the war in Ukraine on cryptocurrencies like Bitcoin and Ethereum between February 24 to June 21, 2022. The results obtained show a negative impact in most cases and point out that decision makers should be more concerned by the asymmetric connectedness between markets. While not related, yet in the same area of economic implications, Wiertz, Kuhn & Mattissek (2023) analyze one year of Twitter, newspapers, and political talks discourses regarding energy transition between February 24, 2021 – February 24, 2022. The study reports the following four discursive shifts caused by geopolitical turn: controversies regarding gas import, new moral imperative to reduce dependence on Russian gas, increase in popularity of resources like nuclear power, accelerated implementation through funding for renewable energy solutions.

Kravchenko (2022) investigates anti-Ukrainian discourses of Russian politicians with a high degree of granularity focusing on topics like the liberation of Ukraine, the Villian role, the hero-liberator construction, and the we-they dichotomy. Furthermore, Hanley, Kumar & Durumeric (2023) conduct a comprehensive discourse analysis on Western, Russian, and Chinese discourses carried out between January 01 up to April 15, 2022. The analysis focuses on three news ecosystems: Bucha, Mariupol, and general events. Accordingly, the study mentions the difference in perspective between the three types of discourses. While on the Western side the focus is more on military aspects and destruction (e.g., *dead, killed, bodies*), the Russian side has the focus on Donbas region and the portrayal of the Ukrainian government (e.g., *repel Kiev's aggression, liberate the regions*), the Chinese side focuses on its concerns regarding economic and diplomatic fallout of the war in Ukraine (e.g., e.g., *negotiations*). Agarwal, Punn & Sonbhadra (2022) analyze public discourses and their dynamics on Twitter using the Kaggle Russia-Ukraine war dataset, with data between December 31, 2021, to March 03, 2022. Moreover, the study depicts the word cloud representation, a summary of tweets and likes, and the frequency of positive, negative, and neutral tweets. Among the topics mentioned are *standwithukrinae, Ukraine border, troops, border, Russia, NATO, Ukraine crisis,* and *people.* Moreover, Haq et al. (2022) build a Twitter dataset between the start of the war until March 06, 2022. The dataset contains 1.6 million tweets, and among the topics with the highest frequency found are *breaking, news, exclude, amid,* and *suspends*. Shevtsov et al. (2022) provide a Twitter dataset from the beginning of the war up to April 07, 2022, containing 57.3 million tweets. The study mention English, French, German, Spanish, and Italian among the most popular languages of the tweets collected. Hakimov & Cheema (2023) advance a Twitter dataset with approximately 1.5 million tweets in 60 different languages with tweets between February 01, 2022, to May 31,

2023. The top languages rated are English, Arabic, Spanish, French, and German. Moreover, a general negative sentiment can be seen with slight improvements in moments of time such as the Independence Day in August 2022 and the visit of Biden in Kyiv in March 2023. Although on a different platform, these results are aligned with the results obtained in our article.

In particular, Caprolu, Sadighian & Di Petro (2022) conduct a Twitter aspect-based sentiment analysis between January 27 up to March 23, 2022, focusing on five types of accounts: baby, trusted, abnormal, unknown, and regular. The study reveals that celebrities and popular accounts were from the beginning active in posting and engaging in different discourses related to this topic. Furthermore, Chen & Ferrara (2023) propose a Twitter dataset between February 22, 2022, to January 08, 2023, and analyze both the discourses and relations between the users engaged. The top five languages are English, Spanish, French, German, and Italian, and among the correspondences between country of users can be mentioned US -> US, Ukraine -> US, US -> Japan, for the pair quoted country -> quoter country. For tackling different social media manipulation, Toraman et al. (2022) analyze misinformation tweets in English and Turkish for building a deep learning-based misinformation detection on Russia-Ukraine war, Covid-19, and refuges. The data collected in English is between September 10, 2020, to March 21, 2022, and the data collected in Turkish is between October 5, 2020, to March 11, 2022. Using a similar approach, Thapa et al. (2022) build a multi-modal Twitter dataset (i.e., both text and image data) with data from February 22 to March 28, 2022. The authors build Generative AI models using techniques like RoBERTa and CLIP for content classification as being hate speech or not related, obtaining values of F1 score of 0.83 and 0.89, respectively. Similarly, Pierri et al. (2022) bring a Facebook and Twitter dataset of 250 million tweets with data ranging from February 01 to April 24, 2022, and propose a propagation and misinformation detection solution.

On other less classical platforms, Fung & Ji (2022) propose the Weibo dataset with 27.341 posts captured from the beginning of the war up to March, 08, 2022. The top five topics discussed are *Ukraine, Russia, Russia-Ukraine, Kiev,* and *Putin.* Park et al. (2022) build a dataset with more than Twitter and VKontacte 38 million messages between February 24 to July 31, 2022, from state-affiliated and independent media outlets. The authors frame the discourses tackled by both communities and position topics regarding *morality, fairness, legality,* and *crime* in the independent sphere, and topics regarding *capacity* and *external* in the state-affiliated sphere. Zhu et al. (2022) propose a Reddit dataset containing more than 300.000 posts between February 24 to May 29, 2022. The study classifies the content as being related to this conflict or general military related. The authors state that a general military post is more likely to receive more replies than the Russian-Ukraine conflict. On Tiktok, Steel, Parker & Ruths (2023) build a dataset around the war in Ukraine containing videos and comments from 2022. The top five languages identified are German, English, Indonesian, Russian, and Ukrainian, and the top five leader mentions are Putin, Zelensky, Biden, Trump, and Macron.

Hanley & Durumeric (2023) conduct an in-depth analysis on communication behavior and topic analysis of Russian media outlets and Telegram channels between January 01 to September 22, 2022. The authors stress that with time, during war, a decrease of using Western applications by Russian users was seen in parallel with an increase in using Telegram channels. Among the topics discussed are the possibility of achieving peace in Donbas in Telegram specific discourses, cessation of hostilities and political dialogue in Russian site specific discourses, and advancements of airborne units in shared Telegram and Russian site discourses. Nazaruk (2022) takes an archiving perspective by collecting 4 TB of Telegram data from around 1000 channels with communication between February 24 to July 31, 2022. The channels are grouped in categories like official and news channels, IT community/information and cyber-attacks, border crossings, Russian propaganda and fakes, medical and psychological assistance, and military mobilization. Ptaszek, Yuskiv & Khomych (2023) analyze Telegram communication between February 24 to June 03, 2022, on RIAN channel which is controlled by the Russian government and UNIAN which is the first and largest independent news agency from Ukraine. While the most used words on RIAN are *Russia, Ukraine, military, military of defence, Putin, DPR, USA, head, sanctions, foreign minister, troops, forces, Mariupol, operations, Zelensky, negotiations, residents,* and *city*, on UNIAN the most used words are *Ukraine, Russia, occupants, invasion, situation, solution, the border, armed forces, military, action, forces,* and *destroy.*

Table 1: Related studies and proposed datasets

No	Objective	Dataset period	Source
1	Economic sanctions analysis	2014 – 2018	Sedrakyan (2022)
2	PTSD search analysis	January – March 2022	Alibudbud (2022)

No	Objective	Dataset period	Source
3	Cryptocurrency impact analysis	February 24 – June 21, 2022	Khalfaoui, Gozgor & Goodell (2023)
4	German energy discourses shift analysis	February 24, 2021 – February 24, 2022	Wiertz, Kuhn & Mattissek (2023)
5	Anti-Ukrainian discourse analysis of Russian politicians	Not acknowledged, study published in 2022	Kravchenko (2022)
6	Western, Russian, and Chinese media outlets discourse analysis	January 01 – April 15, 2022	Hanley, Kumar & Durumeric (2023)
7	Public opinion analysis on Twitter	December 31, 2021 – March 03, 2022	Agarwal, Punn & Sonbhadra (2022)
8	Twitter dataset	February 24 – March 06, 2022	Haq et al. (2022)
9	Twitter dataset	February 24 – April 07, 2022	Shevtsov et al. (2022)
10	Twitter dataset and sentiment analysis	January 27 – March 23, 2022	Caprolu, Sadighian & Di Petro (2022)
11	Twitter English and Turkish dataset, war, and refuges misinformation detection	September 10, 2020 – March 21, 2022, for English, and October 5, 2020 – March 11, 2022, for Turkish	Toraman et al. (2022)
12	Twitter and VKontakte dataset, state-affiliated and independent discourse analysis	February 24 – July 31, 2022	Park et al. (2022)
13	Twitter dataset	February 01, 2022 – May 31, 2023	Hakimov & Cheema (2023)
14	Twitter dataset	February 22, 2022 – January 8, 2023	Chen & Ferrara (2023)
15	Multi-modal Twitter dataset, hate speech detection	February 22 – March 28, 2022	Thapa et al. (2022)
16	Facebook and Twitter dataset, propaganda, and misinformation detection	January 01 – April 24, 2022	Pierri et al. (2022)
17	Weibo dataset	February 24 – March 08, 2022	Fung & Ji (2022)
18	Reddit dataset and discourse analysis	February 24 – May 29, 2022	Zhu et al. (2022)
19	TikTok dataset and discourse analysis	February 24 – December 31, 2022	Steel, Parker & Ruths (2023)
20	Russian media outlets and Telegram behaviour and topic analysis	January 01 – September 22, 2022	Hanley & Durumeric (2023)
21	Telegram dataset and archiving	February 24 – July 31, 2022	Nazaruk (2022)
22	Telegram news agency discourse analysis	February 24 – June 03, 2022	Ptaszek, Yuskiv & Khomych (2023)

The extensive literature review conducted in this research and summarized din Table 1, reveals two important aspects regarding existing solutions built for understanding discourses, topics, and sentiments of users. First, illustrates the complex nature of this phenomenon and the diversity of aspects tackled. And second, shows that while most studies focus on traditional platforms like Twitter, there is a need for more research and AI-based solutions for understanding the same discourses, topics, and emotions also in non-conventional platforms like Telegram and TikTok. This represents the knowledge gap that this research tackles building on previous work in this domain for encouraging further efforts in this direction.

3. Research Approach and Implementation

This research aims to capture topics and corresponding sentiments in Telegram discourses of Ukrainian users building upon previous work (Maathuis & Kerkhof, 2023a; Maathuis & Kerkhof, 2023b) by developing a series of Machine Learning models in a transdisciplinary approach. Accordingly, the Design Science Research Methodology is followed in a Data Science approach (Shearer, 2000; Peffers et al., 2007; Chockalingam & Maathuis, 2022b) and the following research activities are taken:

Goal definition: although during war, Telegram became increasingly used by Ukrainian and Russian communities, still a limited number of studies in the existing body of knowledge focus on Ukrainian

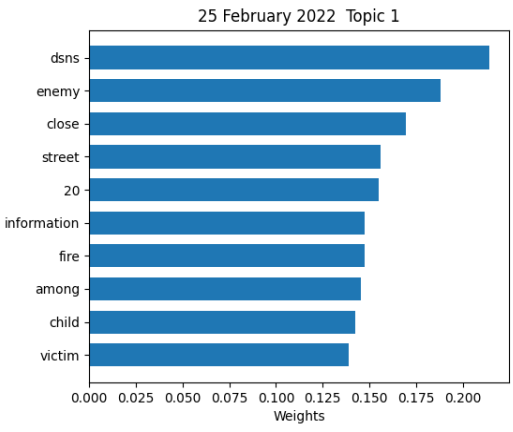
users/speakers in more unconventional social media platforms like Telegram. Hence, this research aims to analyse the discourses, capture topics tackled, and their corresponding sentiments based on data collected in the first year of war, i.e., from February 24, 2022, until February 23, 2023.

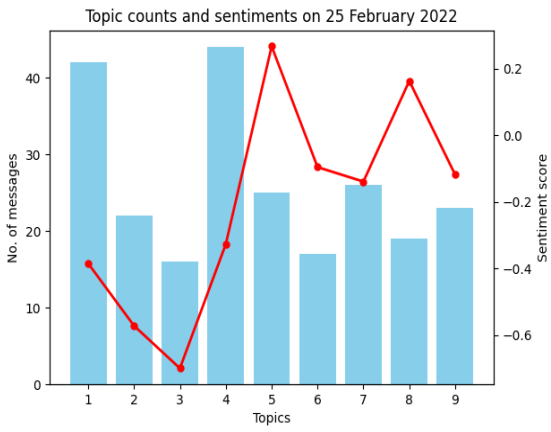
Solution development: the dataset contains 47483 messages in Ukrainian language which are translated using the Google Translate API, and further processed. This implies cleaning the data to remove aspects like unwanted duplicates, stop words, and punctuation. After performing PoS (Parts of Speech) tagging for gathering the grammatical function of a word in a sentence and lemmatization for capturing the semantic root of a word (lemma), the ML models are built for analysing the topics and sentiments using corresponding techniques. In topic modelling, the hidden thematic structure is captured by identifying and extracting topics as clusters of themes tackled in text while facilitating understanding the key structure of the discourses being analyzed (Zhao et al., 2021). Herein, the NMF (Non-negative Matrix Factorization) with Kullback-Leibler Divergence algorithm is applied for factorizing a document-term matrix that contains rows with documents, columns with terms, and cell values with term frequencies within documents. The matrix is split in two parts corresponding to document-topic and topic-term, fact that contributes to easily interpret the results obtained (Sokolova et al., 2016; Hien & Gillis, 2021; Maathuis & Kerkhof, 2023c). In sentiment analysis, the underlying emotional tone and emotions of users are captured through classification (Ates, E. C., Bostanci, E., & Guzel, 2021). In this research, the sentiments are classified as being negative or positive in respect to the topic and per day making use of pre-trained classifier Flair (Magajna, 2022; Maathuis & Godschalk, 2023).

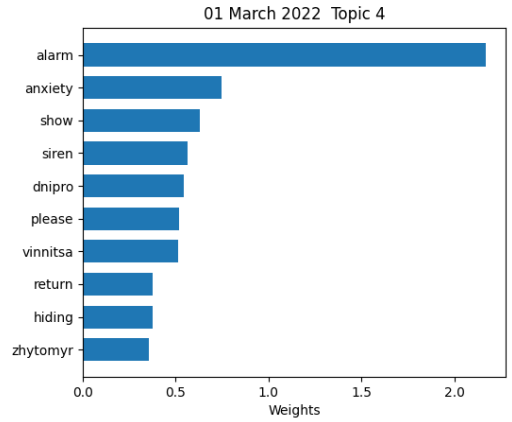
Evaluation and communication: the results of this research are positioned in the existing body of knowledge and ongoing public discourses in a transparent way (Maathuis & Chockalingam, 2022a), and the corresponding findings are communicated through scientific presentations and present article.

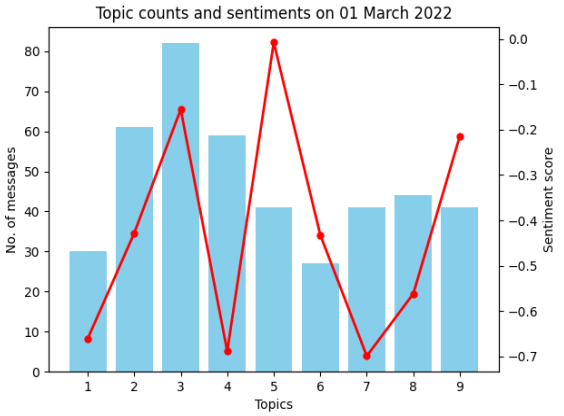
4. Results

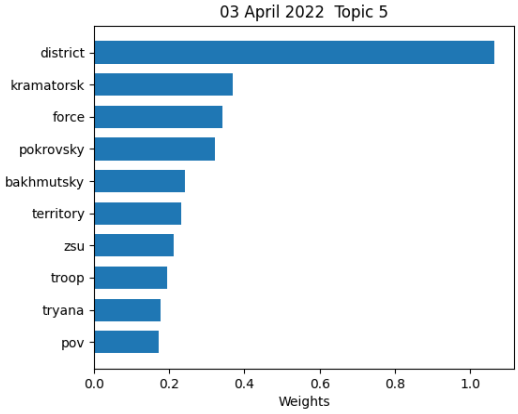
The results obtained from applying topic modelling and sentiment analysis are further presented. On this behalf, a selection of the results are structured in three categories. The first category captures on the left side a selection of nine topics discussed per day and on the right side counts topics and corresponding sentiments per day. For the whole year of study, a day was randomly selected to illustrate both topics and sentiments in Figure 1. Here, the topic numbers found at the bottom of the right image correspond to the topic number at the top of the left image. For instance, topic 1 has a negative sentiment score of -0.4. In the second category illustrated in Figure 2 are depicted words characterizing a topic together with their computed sentiment scores where a score of -1 means negative and +1 means positive for the whole time interval. And in the third category shown in Figure 3 are presented the sentiment scores per day for the whole year analysis.

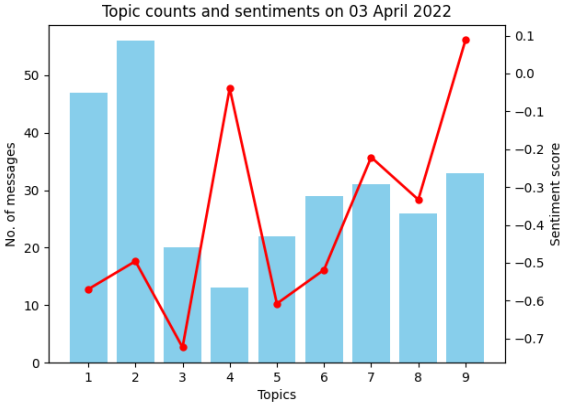


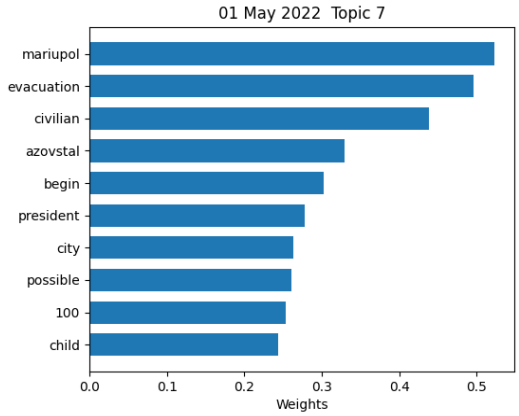


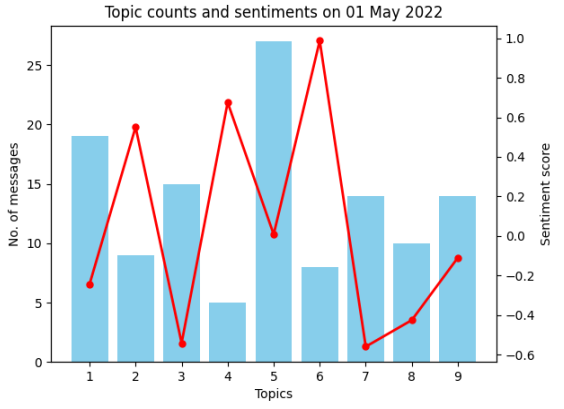

01 March 2022 Topic 4
alarm
anxiety
show
siren
dnipro
please
vinnitsa
return
hiding
zhytomyr
0.0
0.5
1.0
1.5
2.0
Weights

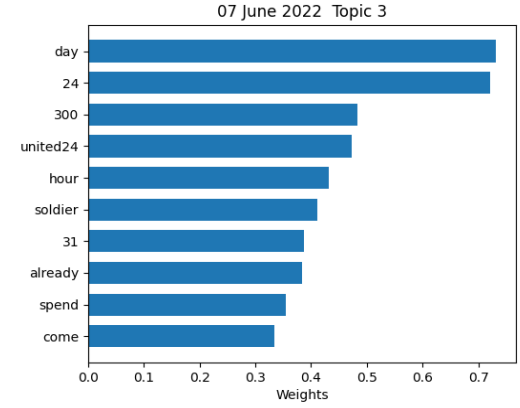

Topic counts and sentiments on 01 March 2022
No. of messages
Sentiment score
Topics

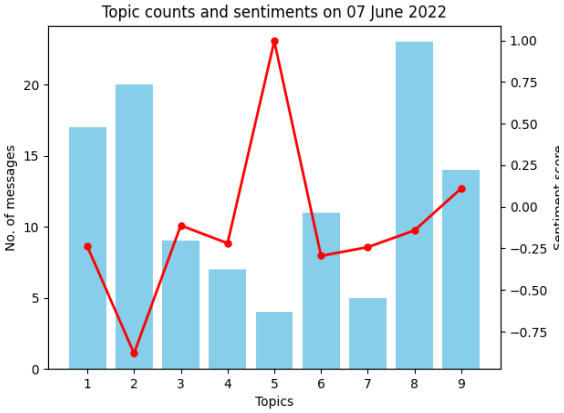

03 April 2022 Topic 5
district
kramatorsk
force
pokrovsky
bakhmutsky
territory
zsu
troop
tryana
pov
0.0
0.2
0.4
0.6
0.8
1.0
Weights

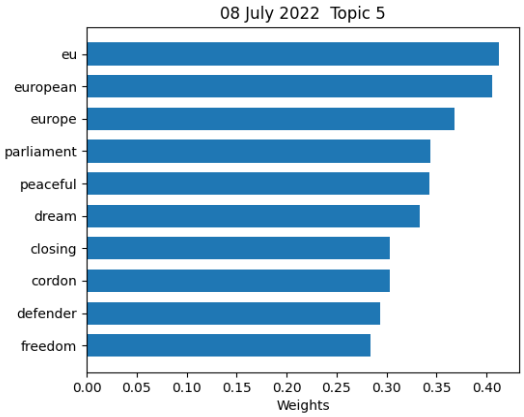

Topic counts and sentiments on 03 April 2022
No. of messages
Sentiment score
Topics

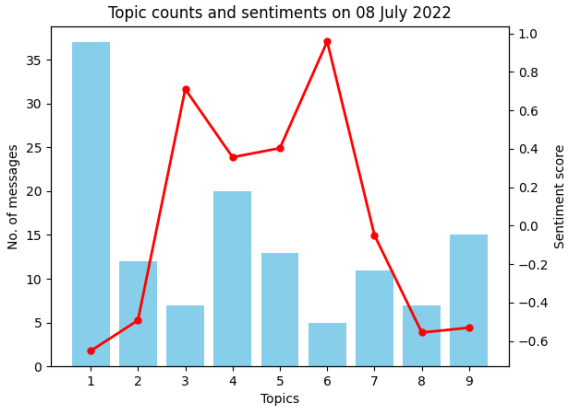

01 May 2022 Topic 7
mariupol
evacuation
civilian
azovstal
begin
president
city
possible
100
child
0.0
0.1
0.2
0.3
0.4
0.5
Weights

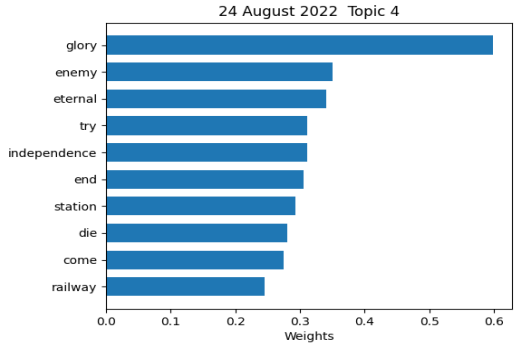

Topic counts and sentiments on 01 May 2022
No. of messages
Sentiment score
Topics

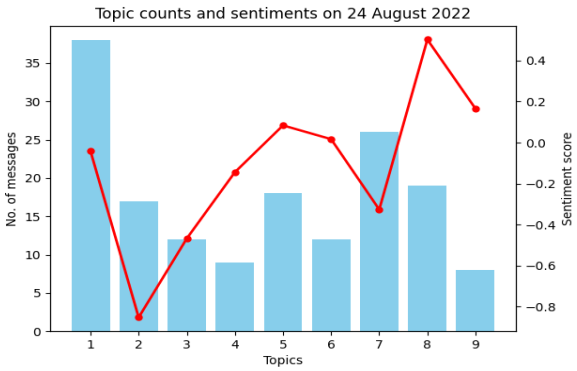

07 June 2022 Topic 3
day
24
300
united24
hour
soldier
31
already
spend
come
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Weights

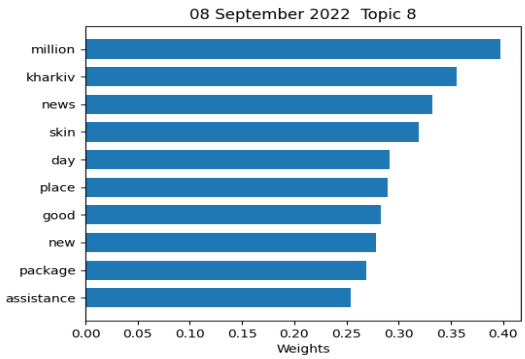

Topic counts and sentiments on 07 June 2022
No. of messages
Topics

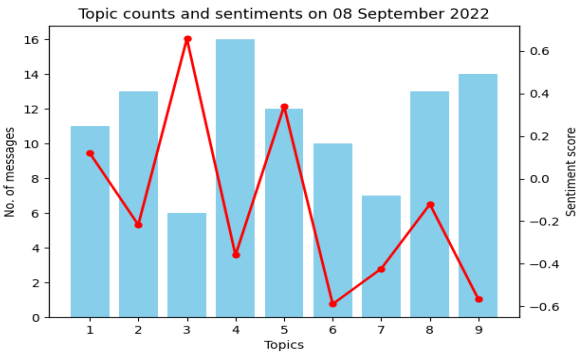

08 July 2022 Topic 5
eu
european
europe
parliament
peaceful
dream
closing
cordon
defender
freedom
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Weights

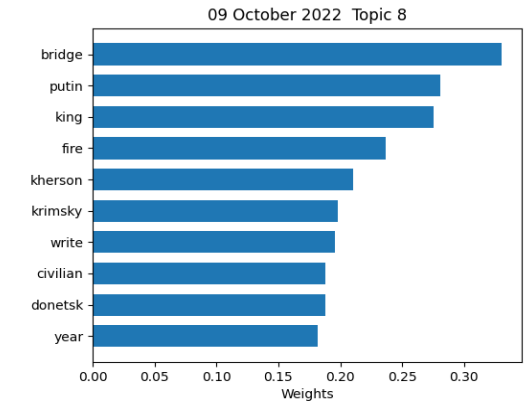

Topic counts and sentiments on 08 July 2022
No. of messages
Sentiment score
Topics

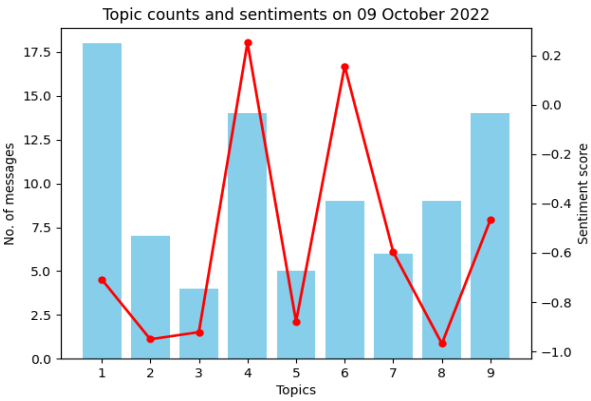

24 August 2022 Topic 4
glory
enemy
eternal
try
independence
end
station
die
come
railway
0.0 0.1 0.2 0.3 0.4 0.5 0.6
Weights


Topic counts and sentiments on 24 August 2022
No. of messages
Sentiment score
Topics


08 September 2022 Topic 8
million
kharkiv
news
skin
day
place
good
new
package
assistance
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40
Weights


Topic counts and sentiments on 08 September 2022
No. of messages
Sentiment score
Topics


09 October 2022 Topic 8
bridge
putin
king
fire
kherson
krimsky
write
civilian
donetsk
year
0.00 0.05 0.10 0.15 0.20 0.25 0.30
Weights


Topic counts and sentiments on 09 October 2022
No. of messages
Sentiment score
Topics

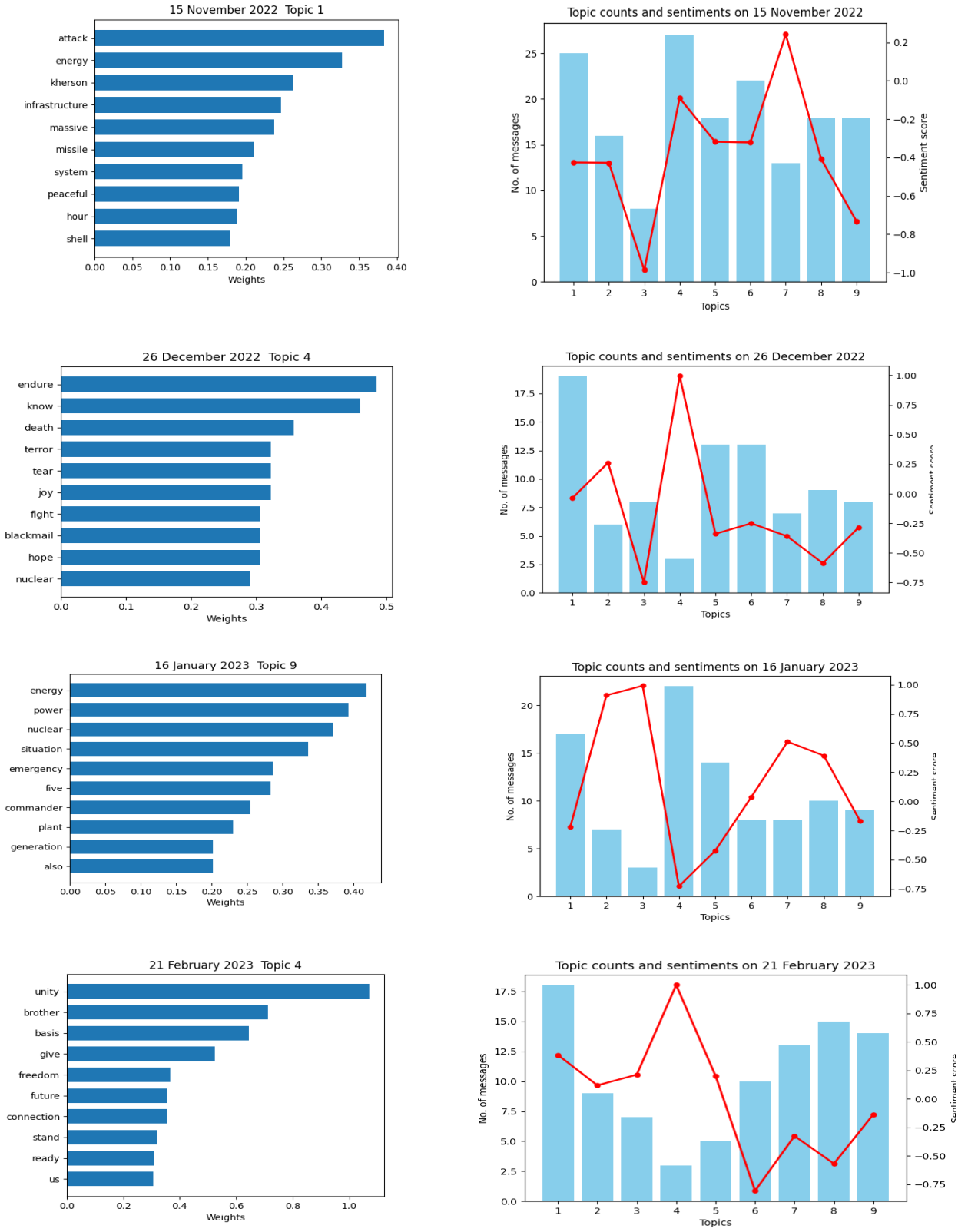


Figure 1: Topics (left side) and topic counts (right side)

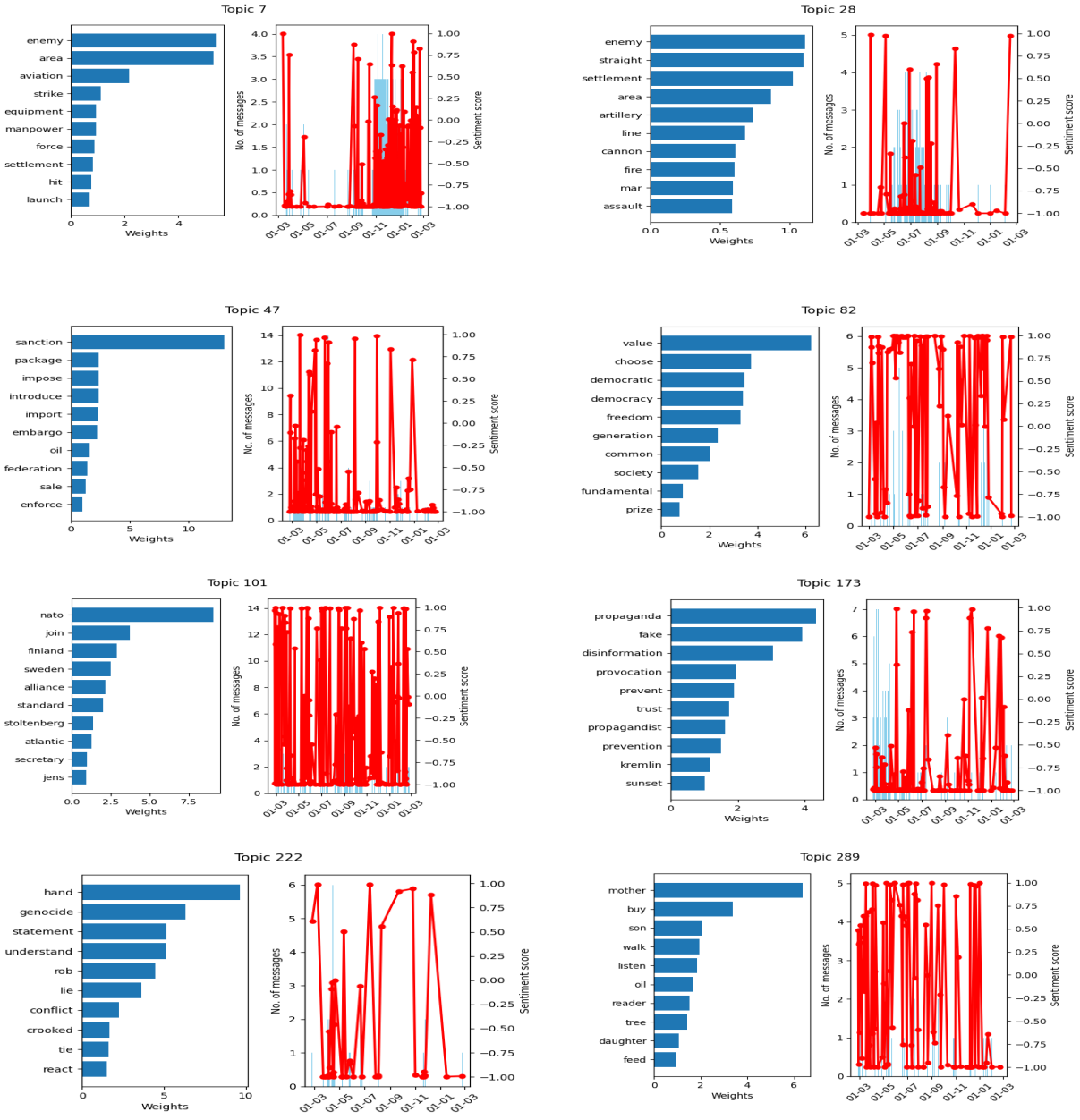


Figure 2: Words for topic with corresponding sentiment scores

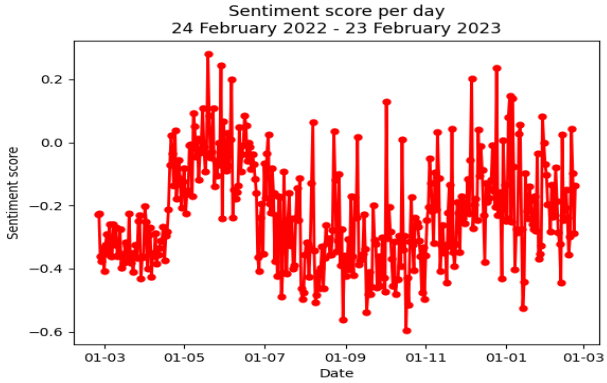


Figure 3: Sentiment scores per day for the whole year analysis

From these results, topics like enemy, victim, alarm, anxiety, territory, troop, Mariupol, president, child, evacuation, soldier, European, freedom, glory, independence, Kharkiv, assistance, civilian, Donetsk, bridge,

Kherson, energy, missile, terror, nuclear, brother, and future, were identified. These topics match public and media discourses plus research and practitioner studies about the war in Ukraine since they capture war-related aspects, human emotions, and names of major cities that were severely affected (UN Secretary General, 2022; NATO, 2022; Levy & Leaning, 2022; Clark, 2023; NATO, 2023), and contribute to countering disinformation and misinformation efforts that aim at altering the public opinion regarding the events and impact characterizing this war and further creating confusion. Moreover, the general sentiment tendency is negative, but is explicable accounting the gravity of the events. Nevertheless, in some moments that have a direct real and more positive meaning, e.g., the Independence Day in Ukraine in August 2022 or the first visit of Biden in Kyiv in February 2023, a slight general tendency of higher sentiment scores is seen. This can be explained by the strength, resilience, and hope that Ukrainian people show in this trajectory.

5. Conclusions

The enduring Ukraine war has attracted substantial global media attention, became a crucial topic in strategic discourses and policies, and through different research and practitioner efforts, various and unique of its political, historical, and cultural angles were analyzed in diverse contexts (Hakimov & Cheema, 2023). Moreover, far-reaching implications like reshaping the power among nations, humanitarian aid, migration, and crisis environmental and climate policies e.g., energy transition programs are increasingly experienced (Wiertz, Kuhn & Mattissek, 2023). It is a complex conflict whose effects extend well beyond the battlefield, who captured an important segment of social media content (Agarwal, Punn & Sonbhadra, 2022) through topics like attacks, separation of families, lack of resources, poverty, and its global impact on multiple fronts (Lava et al., 2022). Nevertheless, while a large number of studies focused on platforms like Twitter and Facebook, less datasets are available and limited attention was given to studying discourses and feelings from unconventional and increasingly used platforms like Telegram and Reddit. On this behalf, this research tackles the existing knowledge gap of focusing exclusively on Ukrainian content about the war. It does that by conducting an in-depth analysis of discourses and corresponding sentiments of Ukrainian Telegram users by building ML models using the Design Science Research methodology in a Data Science approach.

The analysis is conducted on the first year of war, thus between February 24, 2022, to February 23, 2023. The results show a direct match between the discourses carried out in this platform and general media discourses by capturing both human and war-related topics. Additionally, taking into consideration the gravity of the events, a general negative tendency is noticed in the sentiments corresponding to the topics discussed. Further, this research directly contributes to ongoing efforts regarding building transparent and responsible AI-solutions for understanding the dynamics involved in the ongoing war in Ukraine and to supporting the efforts oriented to aborting or reducing social media manipulation mechanisms like disinformation and misinformation about the events, organizations, and people involved in this conflict. For further research, a call to building social media security awareness in time of conflict by developing transparent and responsible AI and gaming solutions is made. This can be done through a transdisciplinary research approach where efforts would be joined for building a safe, responsible, robust, and trustworthy digital domain.

References

Agarwal, N. S., Punn, N. S., & Sonbhadra, S. K. (2022). Exploring Public Opinion Dynamics on the Verge of World War III Using Russia-Ukraine War-Tweets Dataset. *KDD-UC.*

Alibudbud, R. (2022). A web-based analysis of public concern about Post-Traumatic Stress Disorder (PTSD) Ukraine during the 2022 military conflict: An infodemiological study. *European Journal of Trauma & Dissociation*, vol. *6*, no. 3, 100286.

Ates, E. C., Bostanci, E., & Guzel, M. S. (2021). Machine Translation, Sentiment Analysis, Text Similarity, Topic Modelling, and Tweets: Understanding Social Media Usage Among Police and Gendarmerie Organizations. *arXiv preprint arXiv:2101.12717*.

Caprolu, M., Sadighian, A., & Di Petro, R. (2022). Characterizing the 2022 Russo-Ukrainian Conflict through the Lenses of Aspect-Based Sentiment Analysis: Dataset, Methodology, and Preliminary Findings. *arXiv preprint arXiv:*2208.04903.

Chen, E., & Ferrara, E. (2023). Tweets in time of conflict: A public dataset tracking the twitter discourse on the war between Ukraine and Russia. In *Proceedings of the International AAAI Conference on Web and social media*, vol. 17, pp. 1006-1013.

Clark, J. (2023). Biden meets with Ukraine President, as US sends $325M arms package. https://www.defense.gov/News/News-Stories/Article/Article/3534306/biden-meets-with-ukraine-president-as-us-sends-325m-arms-package/

Fard, A. E., & Maathuis, C. (2021). Toward Capturing the Underlying Offensive Mechanisms of Social Manipulation: A Data Model Approach.

Hanley, H. W., Kumar, D., & Durumeric, Z. (2023). " A Special Operation": A Quantitative Approach to Dissecting and Comparing Different Media Ecosystems' Coverage of the Russo-Ukrainian War. In *Proceedings of the International AAAI Conference on Web and social media,* vol. 17, pp. 339-350.

Fung, Y. R., & Ji, H. (2022). A weibo dataset for the 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.05967*.

Hakimov, S., & Cheema, G. S. (2023). Unveiling Global Narratives: A Multilingual Twitter Dataset of News Media on the Russo-Ukrainian Conflict. *arXiv preprint arXiv:2306.12886*.

Haq, E. U., Tyson, G., Lee, L. H., Braud, T., & Hui, P. (2022). Twitter dataset for 2022 russo-ukrainian crisis. *arXiv preprint arXiv:2203.02955*.

Hanley, H. W., & Durumeric, Z. (2023). Partial Mobilization: Tracking Multilingual Information Flows Amongst Russian Media Outlets and Telegram. *arXiv preprint arXiv:2301.10856*.

Hien, L. T. K., & Gillis, N. (2021). Algorithms for nonnegative matrix factorization with the Kullback–Leibler divergence. *Journal of Scientific Computing*, vol. *87*, no. 3, pp. 1-32.

Khalfaoui, R., Gozgor, G., & Goodell, J. W. (2023). Impact of Russia-Ukraine war attention on cryptocurrency: Evidence from quantile dependence analysis. *Finance Research Letters*, *52*, 103365.

Kravchenko, N. (2022). Manipulative Argumentation in Anti-Ukrainian Discourse of Russian Politicians: Integration of Discourse-Analytical and Classical Rhetorical Approaches. *Cogito: Multidisciplinary Res. J.*, *14*, 224.

Lava, S. A., de Luca, D., Milani, G. P., Leroy, P., Ritz, N., & de Winter, P. (2022). Please stop the Russian-Ukrainian war–children will be more than grateful. *European Journal of Pediatrics*, vol., *181*, pp. 2183-2185.

Levy, B. S., & Leaning, J. (2022). Russia's War in Ukraine—The Devastation of Health and Human Rights. *New England Journal of Medicine*, vol. *387*, no. 2, pp. 102-105.

Maathuis, C., & Chockalingam, S. (2022a). Responsible digital security behaviour: Definition and assessment model. In *European Conference on Cyber Warfare and Security* (Vol. 21, No. 1).

Chockalingam, S., & Maathuis, C. (2022b). An ontology for effective security incident management. In *International Conference on Cyber Warfare and Security* (Vol. 17, No. 1, pp. 26-35). Academic Conferences International Limited.

Maathuis, C., & Kerkhof, I. (2023a). The first two months in the war in Ukraine through topic modeling and sentiment analysis. *Regional Science Policy & Practice*, *15*(1), 56-74.

Maathuis, C., & Kerkhof, I. (2023b). First Six Months of War from Ukrainian topic and sentiment analysis. In *European Conference on social media* (Vol. 10, No. 1, pp. 163-173).

Maathuis, C., & Kerkhof, I. (2023c). Social Media Manipulation Awareness through Deep Learning based Disinformation Generation. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 227-236).

Maathuis, C., & Godschalk, R. (2023). Social Media Manipulation Deep Learning based Disinformation Detection. In *International Conference on Cyber Warfare and Security* (Vol. 18, No. 1, pp. 237-245).

Magajna, T. (2022). Natural Language Processing with Flair: a practical guide to understanding and solving NLP problems with Flair. Packt Publishing.

Maschmeyer, L., Abrahams, A., Pomerantsev, P., & Yermolenko, V. (2023). Donetsk don't tell–'hybrid war'in Ukraine and the limits of social media influence operations. *Journal of Information Technology & Politics*, pp. 1-16.

Nazaruk, T. (2022). Subscribe and follow. telegram and responsive archiving the war in ukraine. *Sociologica*, vol. *16*, no. 2, pp. 217-226.

NATO (2022). NATO allies condemn Russia's invasion in Ukraine in the strongest possible terms. https://www.nato.int/cps/en/natohq/news_192406.htm?selectedLocale=en

NATO (2023). Vilnius Summit Communique. https://www.nato.int/cps/en/natohq/official_texts_217320.htm

Park, C. Y., Mendelsohn, J., Field, A., & Tsvetkov, Y. (2022). Challenges and opportunities in information manipulation detection: An examination of wartime Russian media. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 5209-5235.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, vol. *24*, no. 3, pp. 45-77.

Pierri, F., Luceri, L., Jindal, N., & Ferrara, E. (2023). Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. In *Proceedings of the 15th ACM Web Science Conference 2023*, pp. 65-74.

Ptaszek, G., Yuskiv, B., & Khomych, S. (2023). War on frames: Text mining of conflict in Russian and Ukrainian news agency coverage on Telegram during the Russian invasion of Ukraine in 2022. *Media, War & Conflict*, 17506352231166327.

Sedrakyan, G. S. (2022). Ukraine war-induced sanctions against Russia: Consequences on transition economies. *Journal of Policy Modeling*, vol. *44*, no. 5, pp. 863-885.

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *Journal of data warehousing*, pp. 13-22.

Shevtsov, A., Tzagkarakis, C., Antonakaki, D., Pratikakis, P., & Ioannidis, S. (2022). Twitter Dataset on the Russo-Ukrainian War. *arXiv preprint arXiv:2204.08530*.

Sokolova, M., Huang, K., Matwin, S., Ramisch, J., Sazonova, V., Black, R., ... & Sambuli, N. (2016). Topic modelling and event identification from Twitter textual data. *arXiv preprint arXiv:1608.02519*.

Steel, B., Parker, S., & Ruths, D. (2023). Invasion of Ukraine Discourse on TikTok Dataset. *arXiv preprint arXiv:2301.08305*.

Thapa, S., Shah, A., Jafri, F. A., Naseem, U., & Razzak, I. (2022). A multi-modal dataset for hate speech detection on social media: Case-study of russia-ukraine conflict. In *CASE 2022-5th Workshop on Challenges and Applications of Automated Extraction of Socio-Political Events from Text, Proceedings of the Workshop*. Association for Computational Linguistics.

Toraman, C., Ozcelik, O., Şahinuç, F., & Can, F. (2022). Not Good Times for Lies: Misinformation Detection on the Russia-Ukraine War, COVID-19, and Refugees. *arXiv preprint arXiv:2210.05401*.

United Nations Secretary General (2022). Statement by the Secretary General-on Ukraine. https://www.un.org/sg/en/content/sg/statement/2022-02-24/statement-the-secretary-general-%E2%80%93-ukraine

Wiertz, T., Kuhn, L., & Mattissek, A. (2023). A turn to geopolitics: Shifts in the German energy transition discourse in light of Russia's war against Ukraine. *Energy Research & Social Science*, *98*, 103036.

Ye, J., Jindal, N., Pierri, F., & Luceri, L. (2023). Online Networks of Support in Distressed Environments: Solidarity and Mobilization during the Russian Invasion of Ukraine. *arXiv preprint arXiv:2304.04327*.

Zhao, H., Phung, D., Huynh, V., Jin, Y., Du, L., & Buntine, W. (2021). Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.

Zhu, Y., Haq, E. U., Lee, L. H., Tyson, G., & Hui, P. (2022). A reddit dataset for the russo-ukrainian conflict in 2022. *arXiv preprint arXiv:2206.05107*.

Can You Give Me a Lift?

William Mahoney[1] and Adam Spanier[2]
[1]School of Interdisciplinary Informatics, University of Nebraska at Omaha, Nebraska, USA
[2]Cyber Systems Department, University of Nebraska at Kearney, Nebraska, USA

wmahoney@unomaha.edu
aspanier@unomaha.edu

Abstract: Static Analysis (SA) is the practice of examining computer program source code for errors or vulnerabilities outside the compiler's capabilities. To carry out Static Analysis on computer programs, various tools exist that parse and examine each line for known issues. These tools do not compile the program, nor do they run the program. Instead, they analyse the source code directly and infer properties about the program without executing it - thus "static" analysis. Static analysis is not new. Early UNIX contained a program called "lint" for static analysis; a tool that has now existed since the late 1970s (Johnson 1978). Increasingly, however, modern static analysis practices indicate a more focused intent; find cybersecurity weaknesses. At the 22nd European Conference on Cyber Warfare and Security, the authors presented the background of Intermediate Representations (IRs) and described how this "middleware" representation can be utilized for static analysis of source code for cybersecurity weaknesses. By examining an IR, potential flaws in the source code can be located. When utilizing the IR as opposed to the original high-level language, the static analysis process becomes independent of the original source language; if several languages such as C, Rust, and others all compile into the same IR, static analysis of the IR allows the analysis process to no longer be tied to the high-level language grammar or syntax. The previous paper implemented a literature survey of available IR analysis tools to discover prior work; the authors have subsequently advanced the research and are actively using an IR framework, LLVM, for vulnerability analysis. In this research, source code in a high-level language is first compiled to LLVM and the resulting IR is used for analysis. This approach uses a "code-to-IR" SA analysis preparation paradigm. At the same time, there is the potential for binary "lifters" to be used. These tools "lift" an executable program – binary machine instructions – back to LLVM. In this way, the paradigm can also be reversed such that static analysis of LLVM can be performed on source code compiled into LLVM, *or* on executable programs in the field that are "lifted". This begs: how effective are these "lifters"? In this work, the authors present experiences in installation and operation of several binary "lifters" available as open-source projects. Some are supported better than others, some operate better than others, and some don't operate at all. Those that do lead to the follow-up: Is the "lifted" code suitable for static analysis, or is it too obfuscated relative to the original program? This paper describes just that – our efforts and results in locating a binary "lifter" suitable for bringing executable program test cases back to LLVM for analysis by the cybersecurity vulnerability tool concurrently under development.

Keywords: Static analysis, Intermediate representations, Cybersecurity, Vulnerability detection

1. Introduction

Static software analysis (SA) refers to the evaluation of programs for flaws contained within each program's source code (Thompson 2021). Of the two types of analysis, "static" and "dynamic", the latter depends on the execution of the program, while the former examines the code at the source level but does not actually run the program. Generally, static analysis tools operate upon representations of the program by examining the high-level language used to author the software; a program written in C++ would use a static analysis tool specifically for C++, while a program written in Java would necessitate a Java-based analysis tool. Static analysis is not new; UNIX programs used "lint", a small program named after the small bits of fibre in a clothes dryer, as early as 1978 (Johnson 1978, 1979). More recent tools with similar names are still in use today (PC-lint 2023).

The authors have been investigating a novel static analysis approach that implements analysis procedures at the Intermediate Representation (IR) level associated with the compiler rather than at the source code level. For example, the popular Linux compiler suite GCC uses an IR named Gimple (GCC 2023), while the compiler tools based on CLANG use an IR named LLVM (LLVM 2023). As an aside, Gimple is a generic form of "simple" which was a project at McGill University, while LLVM was developed at the University of Illinois and was previously an acronym but is now used stand-alone. The advantage of using the compiler IR as opposed to the high-level code may already be apparent to the reader; if many high-level languages compile into the same IR, analysis at the IR level makes the SA tool independent of the original source code language; what the researchers call *Programming Language Agnostic.* For instance, many languages compile into LLVM, including Fortran, Haskell, Julia, Kotlin, Lua, Objective-C, OpenGL, Rust, and Swift, as well as other oddities including Pony, Crack, and others (LLVM 2023, Pony 2023, Crack 2023). Searching for certain software issues at the LLVM level thus covers all languages that compile into LLVM.

In the past, the authors have reported on research utilizing the LLVM IR in several ICCWS and ECCWS conferences (Mahoney 2021, Mahoney 2022, Spanier 2023). This prior research was aimed primarily at software intellectual

property protection, by “fingerprinting” individual copies of executable programs. The focus in this research has shifted; currently the aim of this project is two-fold. First, the authors are developing a domain-specific language (DSL) with the aim of taking an English software-flaw description and encoding it in an easy, readable language. The resulting description will be used to generate checking code which will examine LLVM for the specified flaw. The English-based descriptions chosen for this research come from the popular Common Weakness Enumeration (CWE) (Mitre 2023). The test cases which will be used for this effort come from the Software Assurance Reference Dataset (SARD) available from the National Institute of Standards and Technology (NIST) (SARD 2023). This research is ongoing, and the authors anticipate reporting on this effort in future publications. A second aim of the ongoing work is to use binary “lifters” to increase the potential coverage for the static analysis tool currently under development. This second aim is the topic of this paper.

What is a “binary lifter”? These are efforts to take an executable program – in binary form as a Windows EXE or a Linux ELF file – and convert the code to LLVM. LLVM, as an IR, is a higher-level representation than the actual program instructions, thus the term “lift”. The process of compiling has a similarity with the process of hashing: during the compiling process, information is lost. Due to this unavoidable loss of information, the process of lifting machine code back to the original high-level representation never results in a high-fidelity representation. Just as there are static and dynamic analysis tools, not executing versus executing the binary, there are static versus dynamic lifters. For this work, either method is acceptable, as the desire is to analyse test cases from SARD regardless of how they are lifted.

Lifting a binary executable back to an IR presents a possibility for this research: one can conceivably start from a Linux/Windows executable file, lift the executable to LLVM, and run static analysis tools on the result. In this way a program can be analysed for potential vulnerabilities, using the CWE data set, whether the original program was compiled *into* LLVM or whether the executable program (with no source code) was *lifted* to LLVM. This possibility theoretically provides a uniform SA tool for all compiled binaries of a given compiler.

The authors thus decided to “test drive” four binary lifting projects and report the results here. Of course, not all are created equal. Since they are open source, some have little to no documentation, some work “right out of the box” and others have serious flaws. This paper reports on the results of downloading, installing, and testing out four open-source projects for binary lifters. As a ramification, this paper presents little in terms of background and prior work, as the prior work is directly represented in the open-source projects described herein. The authors do report individually on the experiences encountered, and for those tools that operate, briefly refer the reader to a paper on performance information. A result is that the paper is organized differently than a traditional research paper. Sections two through five present each of the four lifters, in turn, and overall conclusions are presented in section six.

2. A Failure: McSema

The authors started from a well-known lift tool named McSema, which was presented initially at the Montreal Recon Conference (McSema1 2023, Recon 2014). The McSema team also make available a quite good video of their presentation via YouTube (McSema2 2023) and the source code is in an open repository at GitHub. Hopes are high. From the repository:

> *“McSema is an executable lifter. It translates ("lifts") executable binaries from native machine code to LLVM bitcode. LLVM bitcode is an intermediate representation form of a program that was originally created for the retargetable LLVM compiler, but which is also very useful for performing program analysis methods that would not be possible to perform on an executable binary directly.” (McSema1 2023)*

Although the project is dated, the repository that is openly available does have modifications dated 2021; the authors assume that it is “current enough” that it can be made to work, although possibly requiring Linux and other versions of software from that time period.

But the installation of McSema is, like many open-source projects, fraught with issues. There are two primary problems installing and using this tool. First, the documentation is scant with respect to the necessary required prerequisite packages; some of them are listed in the documentation, but many are not, and apparently the unlisted packages are assumed to be necessary for a correct build of the tool. The documentation for McSema also states that it relies on several other tools which need to be installed first, specifically “remill” and “anvill”. Unfortunately, these projects have similar issues, for example the latter tool requires the package “doctest”, which is not listed as needed for McSema. So, the authors here fall into a loop: find what the build process says

is missing, locate which package for Linux includes that feature, install that package using the Linux installer, start again, discovering the next package. Repeat.

A second problem with McSema is that the documentation that exists on the repository is outdated. For example, the install instructions give very specific releases for the necessary tools, like calling out "release_93aba7c" for "remill", when no such release exists, and showing an example build using LLVM 11 when the only version that seems to make forward progress is version 12, and the current stable version is 15. Additional permissions are needed to copy the files into the correct destination directories, which is a minor issue but is not mentioned in the documentation.

Lastly, McSema looks for a package named "xed" (intel 2023). This is a collection of C code, supplied by intel, which is used to decode x86-64 instructions in software. It is easily downloaded and compiled, and the authors have used "xed" several times with no issues on prior research projects, e.g. (Mahoney 2023). It is not, however, a Linux "package", so despite downloading and compiling "xed", McSema cannot find it as a package, and building the tool stops.

The first attempt at a working binary lifter, McSema, is a failure. However, all is not lost, as the documentation for McSema includes a comparison to other lifter packages, along with links. If McSema is problematic, at least they point in a new direction. Time to try another!

3. Translation by QEMU: Revng

The authors of the open-source tool "revng" have taken a different approach to lifting code. They utilize the CPU emulator often used to host virtual machines (VMs), "QEMU" (Bellard 2005, QEMU 2023). This package, QEMU, is unique in that it allows one to host a VM architecture that is different than the host machine. For example, one can run a Linux virtual machine for the ARM CPU on an intel x86 computer. Another example of cross-CPU emulation is Apple's Rosetta 2 system, allowing a user with x86-64 programs to emulate them on newer Apple computers with their M1 or M2 chips (Evans 2020). Assuming that the processor on the VM is different than on the host, as QEMU encounters instructions they are translated using the Tiny Code Generator (TCG). These individual TCG instructions are then interpreted on the host QEMU system. In this way the "machine" running the VM and the actual machine, running QEMU, need not be the same type of computer; TCG acts as "the middleman" IR. So, "revng" uses the QEMU infrastructure to create TCG, and then to translate the TCG into LLVM. The process starts by handing an executable image to "revng", which uses QEMU to "run" the program. As the internal instructions, in TCG, are encountered, LLVM instructions which match the TCG are emitted.

In the case of this tool the documentation is somewhat better, calling out the prerequisite Linux packages needed, as well as informing the user that they need to install an ancillary tool, "orchestra". This is a command line shell used to set up the proper paths to run the tool. And where to start? With the documented example, which includes the following classical but trivial "Hello World" C test program:

```
#include <stdio.h>
int main( int argc, char *argv[] )
{
   printf("Hello, world!\n");
}
```

Figure 1: Documented test case, from "revng" documentation

After installing "orchestra" and using it to install "revng", as well as a compiler for the ARM architecture, the documentation example compiles the simple program in Figure 1 into an ARM executable program. At this point the executable is translated using the "revng" tool into the native machine (in this case Linux on x86). Next the documentation shows that one can change the permissions on this translated program so that it is executable and run the resulting program natively. With a current Linux version (22.04.1) and following the documentation steps, the resulting translated program simply "core dumps" and exits.

Next the authors noticed an oddity of the translation process. First the "Hello World" program in Figure 1 was compiled twice, once with the GCC compiler and once with the "clang" compiler, into x86-64 binaries. In both cases the programs were compiled with the same options (including "-static") and were not "stripped". The two executable programs are 899 thousand bytes (Kb) and 909Kb, respectively. Compiling the program in Figure 1 into an ARM executable, using the instructions on the repository and the ARM GCC compiler called out in the instructions, generates a file with 123Kb. Lastly, compiling the code on a Raspberry Pi board, which is also ARM and with the same options, gives an executable program that is 517Kb. Note that the x86-64 programs are similar

in size, but the ARM executables are not. The authors did not investigate the reason for this since the second oddity makes it moot.

For the second interesting effect, the authors compiled the program in Figure 1, using "clang", and told the compiler to generate the intermediate LLVM file and stop. The size of this resulting file is 1,781 bytes (not Kb). Following the repository instructions, the ARM executable program was lifted to LLVM. This resulting LLVM file is 4,484Kb, roughly 2,500 times larger.

To attempt to account for this explosion in the size of the resulting translated files, the authors used the Linux tool "objdump" to examine the sections, headers, and disassembly of the contents of the translated executable program. It appears that the result from the "revng" translation is a program designed to interpret the translated IR, not a program translated into actual x86 machine code. For example, there are numerous symbols in the program which begin with "qemu_" such as "qemu_mutex_lock" and others. In addition, functions with names such as "process_pending_signals" implies that the resulting program is designed to run like the original but use the "qemu" infrastructure. Additional clues towards this include many functions with names that start with "target_to_host_" or "host_to_target_". Thus, it appears that the aim is to recreate the same semantics as the original program, but not an actual translation of the original program. This may be incorrect, but it is difficult to tell since the translated program does not run. Time to try another!

4. Success with a Caveat: Retdec

The third binary lifter the authors experimented with is "retdec", the retargetable de-compiler (Avast 2023). The installation instructions for this tool are both complex and simple at the same time; they contain detailed instructions for how to build the programs from scratch on Linux, Windows, MacOS, and FreeBSD. A much easier route is to simply download the compressed Linux "tar" file and extract everything.

The main interface for this tool is a command-line executable program named "retdec-decompiler". The tool has a multitude of command line options, many of which deal with methods for recreating a high-level program code from the original binary. Here the research is to lift a binary to LLVM and stop, but LLVM is an intermediate step in the translation process for this tool. It lifts the binary to LLVM and then converts the LLVM into C code.

An initial test with "retdec" is to take the "Hello World!" program from Figure 1 and run a test that is like the test in the previous section. The extremely simple program is compiled on x86-64 Linux into a Linux executable. The executable is then de-compiled back to C, and a portion of the resulting program is shown in Figure 2:

```
// Address range: 0x1140 - 0x1165
int main(int argc, char ** argv) {
    // 0x1140
    printf("Hello World!\n");
    return 0;
}
```

Figure 2: An initial test of "retdec" decompiler

For trivial code the tool apparently works. However, when decompiling the same program compiled for ARM the results are not quite as great; these results are in Figure 3 and show that the knowledge that the called function is "printf" has been lost somewhere along the line.

```
// Address range: 0x10310 - 0x1031c
int32_t _24_a(void) {
    // 0x10310
    int32_t result; // 0x10310
    return result;
}

// Address range: 0x10430 - 0x10464
int main(int argc, char ** argv) {
    // 0x10430
    _24_a();
    return 0;
}
```

Figure 3: Same program, compiled for ARM and decompiled with "retdec"

When used to recreate a reasonable representation of the original program, at least for x86, the "retdec" tool seems to do a good job and rivals others such as IdaPro and Ghidra in generating appropriate C code (HexRays 2023, Ghidra 2023). For this research the aim is not to decompile back to a high-level language but to lift the binary back to LLVM for potential vulnerability analysis. Thus, the testing of "retdec" in this context is actually concerned, not with the quality of the high-level source code produced, but rather with the quality of the IR created along the way. For this reason no investigation was made into why the ARM code produced was of such low quality.

Similar tests with "retdec" were performed; the simple program was compiled from source code into an LLVM file, and then compiled into an executable which was then lifted back to LLVM. The original compiled file contains 1,423 bytes and the lifted LLVM file contains 4,890 bytes. There is no explosion in size as there was with "revng". Further, a quick look at the reconstructed C source from "retdec" shows why the size is larger. This tool not only attempts to recreate the original program, but also all the ancillary code that is associated with it; this includes the C start-up function "_start" that handles command line parameters, the "_init" and "_fini" functions present in a C/C++ program compiled on Linux, and so on. Figure 4 shows "_init":

```
// Address range: 0x1000 – 0x101b
int64_t _init(void) {
    int64_t result = 0; // 0x1012
    if (*(int64_t *0x3fe8 != 0 ) {
        // 0x1014
        __gmon_start__();
        result = &g3;
    // 0x1016
    return result;
}
```

Figure 4: "retdec" decompiles all executable code, even start-up code

An issue with this "quality" is that one cannot simply take the resulting C program and immediately compile it, as all the extraneous functions will be multiply defined. It's almost "too good" in that sense. Because the ancillary code produced should be the same for each program, the roughly three-times size increase is presumably washed out with a larger executable. To test this hypothesis, the authors selected a slightly more complex program specimen; a program with approximately 200 lines of C code that serves to simulate the functions of a router. When compiled with "clang" to generate LLVM code, the result is 28,550 bytes with 569 lines of IR code. When lifted from the executable the LLVM is 34,226 bytes with 764 lines. A quick examination shows that again this difference is principally the addition of the start-up functions being decompiled as well.

The conclusion of the authors is that with one minor issue, this tool successfully lifts at least x86-64 binaries to LLVM and as a nice side effect generates a quite good representation of the original source code. The one minor issue? The generated IR is for a prior LLVM version and cannot directly be used with a current LLVM package.

5. Microsoft Jumps In: llvm-mctoll

Lastly the authors installed and tested a tool currently available on GitHub called "llvm-mctoll". This tool was written and reported on primarily by Yadavalli (2019) from Microsoft. While the project comes in source code form, the build procedure (unlike many other open-source projects) is largely described in simple terms with minimal mistakes. In fact, the project appears to be very active. As updates to the project are made, the current version of the LLVM software necessary is not only specified in a ".txt" file in the repository but is also a relatively current and supported version of the IR. Additionally it handles the "Hello world" program of Figure 1 with only one caveat, generating a reasonably sized LLVM file as a result.

The caveat mentioned has to do with the call made in Figure 1 to the common C function "printf". Modern CPUs such as x86-64 or ARM call functions and pass the parameters to those functions in CPU registers. It is not always possible to determine what parameters are used at the point of a function call. For instance, the simple "printf" call itself – while it is not shown in the trivial example above – can take a variable number of parameters. Figure 1 only has one, but the function can have an arbitrary number of additional data after the string. It is necessary for a lifting tool to determine which CPU registers are function parameters, as the LLVM "call" instruction requires that these parameters are explicit in the IR. Details of this issue are spelled out by Fink (Fink 2022) but can be summarized; In the process of examining the native instructions, the tool tracks which registers have been assigned to but not changed before the point of the function call. It is significantly more complex than

described here but details are found in the paper by Fink. The caveat referred to above is that this tool requires a user to provide a viable function prototype for any and all functions not known by the tool!

An example of this gotcha is “printf” itself. When lifting a binary, say an x86-64 binary on Linux, it is necessary to tell the tool the format of any call to “printf”. This format is called the function “prototype” and in the case of “printf” is found in a standard file in the compiler infrastructure. In cases where a function is called from the binary code, but the prototype for that function is not known, “llvm-mctoll” stops and will not produce IR at all. All functions must be known by the tool, even if the tool’s user must guess the prototype. Incidentally, the popular reverse engineering tool “ghidra”, developed by NSA, takes a global view across all code in an executable and it propagates the calling convention recovered in any function call across all references to the function (Toor 2022).

The performance of this tool appears to be good and is reported upon in Toor as well. Toor used the Phoenix-2 system, which is an in-memory implementation of the MapReduce data model popularized by Google (Kozyrakis 2023). Unfortunately, this is not a standard benchmark such as SPEC (SPEC 2023) but is fine for reasonable estimates of performance. Details are in the paper but there is reasonably small impact on the size or speed of the lifted (and recompiled) program, assuming that after it is lifted it is optimized by the LLVM tools before being converted back to machine code.

For the practical application desired here, performing vulnerability analysis on LLVM, the question is whether this tool is acceptable. As a test the authors started from one of the SARD vulnerability cases, number 240,080. This file is an example of CWE 415, inadvertently freeing allocated memory twice. The compiled source for this test generates an LLVM file of 272 lines and 10,313 bytes, while the lifted file contains 1,064 lines and 42,922 bytes. This is not an unreasonable expansion. One of the functions within this test case has an example of the vulnerability, and Figure 5 displays the control flow of the bad function from this test case, for the compiled and lifted versions. We conclude that this tool, with supplied function prototypes, will work for our vulnerability scanner.

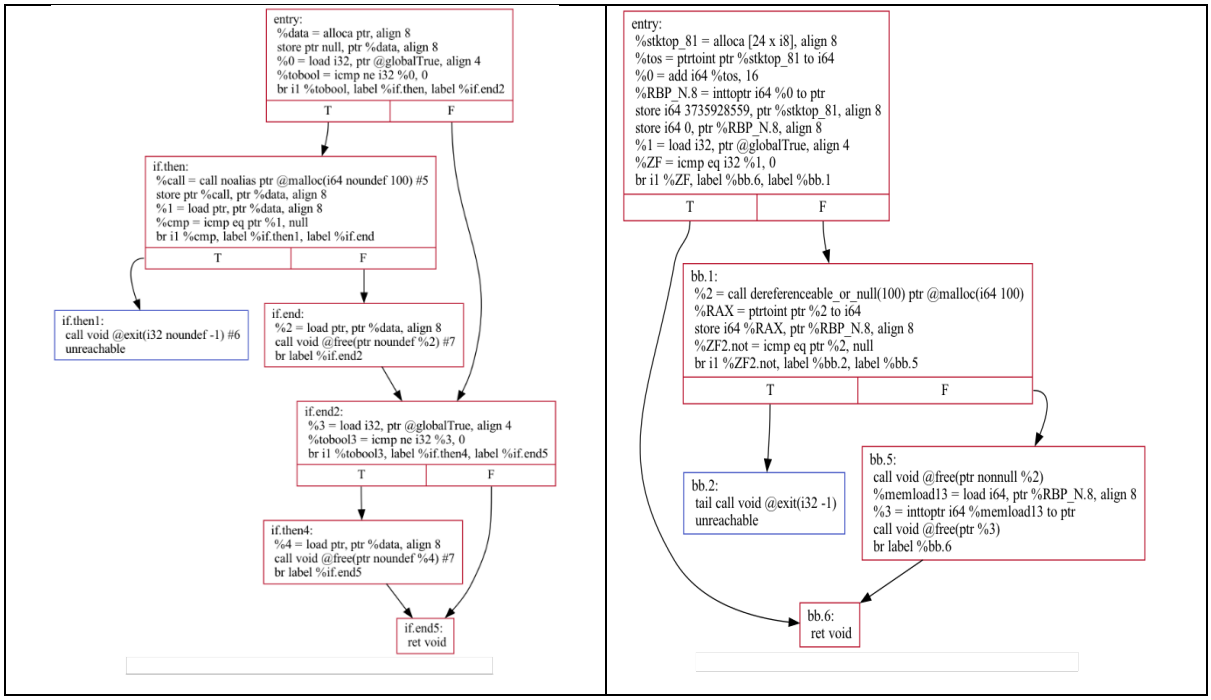


Figure 5: SARD test case “Bad” function compiled (left) and lifted (right)

6. Conclusions and Future Work

Recall that our primary motivation for finding a viable “lifter” for executable programs is the capability to hand the resulting LLVM to the static analysis tools the authors are constructing in a parallel effort. Also recall that the method for testing *that* effort is to compile test cases from the SARD database (SARD 2023). The SARD test cases are intentionally minimal in the sense that they are designed to not require noteworthy external support. Rather, the test cases have a small set of external functions to provide for printing the results of the tests. These few functions, as are the test cases themselves, open for examination. The majority of the test programs needed

for the research effort rely on just a few common files, which contain the necessary prototypes. For example the SARD files “std_testcase.h” and “std_testcase_io.h” have appropriate function headers and thus can simply be includes in the command line for “llvm-mctoll”.

The result is that it is a simple matter to provide function prototypes for the printing capabilities of the test cases. The direct result is that the fourth attempt at locating a “lifter” for our static analysis tool is a successful search; the authors will be proceeding with “llvm-mctoll”.

References

Avast (2023) “RetDec, a retargetable machine-code decompiler based on LLVM”, [online], https://github.com/avast/retdec accessed October 2023.

Crack (2023), [online], available at https://github.com/crack-lang/crack accessed September 2023.

Bellard, F. (2005) “QEMU, a Fast and Portable Dynamic Translator”, Proceedings of the Annual Conference on USENIX (ATEC ’05) , USENIX Association, Berkeley, CA, 41–41.

Evans, J. (2020) “Everything you need to know about Rosetta 2 on Apple Silicon Macs”, [online], Computerworld, November 19, 2020, at https://www.computerworld.com/article/3597949/everything-you-need-to-know-about-rosetta-2-on-apple-silicon-macs.html accessed October 2023.

Fink, Martin (2022) “Translating x86 Binaries to LLVM Intermediate Representation”, Bachelor’s Thesis in Informatics, Technische Universität München, [online], at https://dse.in.tum.de/wp-content/uploads/2022/01/translating_x86_binaries_into_llvm_intermediate_representation.pdf viewed October 2023.

GCC (2023), [online], at https://gcc.gnu.org/onlinedocs/gccint/GIMPLE.html accessed October 2023.

Ghidra (2023), [online], at https://ghidra-sre.org and viewed October 2023.

HexRays IdaPro (2023), [online], http://www.hex-rays.com/products/ida/index.shtml accessed September 2023.

Intel (2023), [online], “Intel XED” at https://intelxed.github.io viewed September 2023.

Johnson, S. C (1978) “Lint, a C Program Checker”, Computer Science Technical Report, October 25 1978, Bell Labs, 78-89.

Johnson, S. C. (1979) “Lint, a Program Checker”, Unix Programmer's Manual, Seventh Edition, Vol. 2B, M. D. McIlroy and B. W. Kernighan, eds. AT&T Bell Laboratories: Murray Hill, NJ, 1979.

Kozyrakis, C. (2023) “Phoenix”, [online], at https://github.com/kozyraki/phoenix accessed October 2023.

LLVM (2023), [online], at https://llvm.org viewed September 2023.

McSema1 (2023), [online], at https://github.com/lifting-bits/mcsema viewed September 2023.

McSema2 (2023), [online], video available at https://www.youtube.com/watch?v=x08mZevfiK8 accessed September 2023.

Mahoney, W., Hoff, G., McDonald, J. T., Grispos, G., (2021) “Software Fingerprinting in LLVM”, 16th International Conference on Cyber Warfare and Security, 2021.

Mahoney, W., Sigillito1, P., Smolinski, J., McDonald, J. T., Grispos (2022) “Analyzing the Performance of Block-Splitting in LLVM Fingerprinting”, 17th International Conference on Cyber Warfare and Security, March 17-18, 2022, Albany, New York, USA.

Mahoney, W., McDonald, J. T., Grispos, G., Mandal, S. (2023) “Improvements on Hiding x86-64 Instructions by Interleaving”, 18th International Conference on Cyber Warfare and Security, March 9-10, 2023, Towson University, Towson, Maryland.

Mitre (2023), “CVE Overview”, [online], at https://www.cve.org/About/Overview viewed September 2023.

PC-lint (2017), [online], at https://pclintplus.com viewed September 2023.

Pony (2023), [online], at https://www.ponylang.io viewed September 2023.

QEMU (2023), [online], at https://www.qemu.org downloaded August 2023.

Recon 2014 (2014), [online], Conference agenda at https://recon.cx/2014/conference.html viewed September 2023

Revng (2023), [online], at https://github.com/revng/revng downloaded August, 2023.

SARD (2023) “NIST Software Assurance Reference Dataset”, [online], at https://samate.nist.gov/SARD/ accessed October 20233.

Spanier, A., Mahoney, W. (2023) “Static Analysis Using Intermediate Representations: A Literature Review”, Proceedings of the 22nd European Conference on Cyber Warfare and Security, Athens Greece, June 22 and 23, 2023, (Digital ISBN 978-1-914587-70-2)

SPEC (2023) “Standard Performance Evaluation Corporation”, [online], at https://www.spec.org/benchmarks.html and accessed October 2023.

Thomson, P. (2021) “Static Analysis: An Introduction. The Fundamental Challenge of Software Engineering is One of Complexity”, ACM Queue, Volume 19, Number 4, 29–41.

Toor, T. S. (2022) “Decompilation of Binaries into LLVM IR for Automated Analysis”, Master of Applied Science in Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada, [online], at https://uwspace.uwaterloo.ca/bitstream/handle/10012/17976/Toor_Tejvinder.pdf?sequence=3&isAllowed=y accessed October 2023.

Yadavalli. S. B., Smith, A. (2019) “Raising binaries to LLVM IR with MCTOLL (WIP paper)”, LCTES 2019: Proceedings of the 20th ACM SIGPLAN/SIGBED International Conference on Languages, Compilers, and Tools for Embedded Systems, June 2019, 213–218.

Cybercrime Classification: A Victimology-Based Approach

Sayonnha Mandal
University of Nebraska Omaha, USA

smandal@unomaha.edu

Abstract: The need for understanding cybercrime and the possibility of its occurrence is significant to mitigate its adverse effects on society. A comprehensive universally agreed-upon classification scheme for cybercrime is hugely lacking in terms of utilizing a complete perspective of the entities involved in the same. A new perspective in cybercrime classification is moving beyond the machines and focusing on the humans, especially the victims. Cyber victimization extends from the single user to a mass or system perspective, thereby representing governments, organizations, and society to be categorized as victims. This paper proposes a novel ontological classification of cyber victimology that can help illustrate the complete cybercrime incident from the perspective of the victim. We utilize a multidimensional typology to represent the dimensions and classifications of the cybercrime victim. We then analyse the semantic relationships between the ontological objects to develop a comprehensive victimology representation. The understanding of the type and role of the victim provides new insight into the analysis of the cyber incident. Moreover, the resultant representation can serve as an extension to current cybercrime ontological frameworks and help in providing a new point of defence in cybercrime incidents. Finally, such a victimology-based classification can subsequently result in a dynamic ontology which can be queried to obtain relevant insights into the nature and occurrence of cybercrimes.

Keywords: Cybercrime, Victimology, Ontological framework, OWL

1. Introduction

In recent years, there has been an increase in cybercrime and consequently individuals, organizations and governments have all been negatively impacted Cybercrime at present is endemic to society at large. Be it in the disruption of business and governments, impacting the lives of people and adversely affecting the economy, cybercrime is an ever-increasing phenomenon today. Moreover, the fact that these can be categorized as global crimes, due to the nature of transcending global boundaries in its impact as well as in its source, makes this a prominent aspect of current cybersecurity research. However, a single or standardized definition of the term "cybercrime" is largely lacking in computer science terminology. Donalds et. (2014) al defines the term as something "generally used to cover/describe a wide variety of illegal crimes or what is considered illicit conduct by individuals/groups against computers, computer-related and other devices, information technology networks, or traditional crimes, as well as actions targeting individuals, supported by the use of the Internet and/or technology." Prominent cyber-attacks such as WannaCry ransomware have global impact, encrypting the theses and graduation projects of students in Chinese universities as well as leading to massive disruptions in medical procedures in Britain. The popular ransomware attack also affected Russia's Health Ministry, Russian Railways and Russian and Spanish telecom companies Megafon and Telefonica respectively. French car company Renault and Brazil's SSN office and FedEx were also some of the bigger organizations adversely affected by WannaCry (Donalds et. al, 2019).

Previous research has focused on the role of the device involved, nature of the attack and the viewpoint of the attacker or the defenders. To categorize and analyse these varied threats and attacks, various legal frameworks have been proposed. These include the Computer Security Incident Response Teams (CSIRTs) that target malicious cyber activities involving the use of information and communication technologies (ICTs). The role of law enforcement and their approach to cybercrime analysis has also been modified to implement knowledge centric policing. It has been identified that at least 30 categories of crime knowledge units are in use for police practice. These are demarcated into administrative, policing, legal, procedural, and analytical. A popular knowledge-based police approach is the Problem Oriented Policing or POP, a term coined by Herman Goldstein, which involves the use of four different types of knowledge to analyse crimes. Intelligence-led policing or I-LP, a more recent improvement on the POP method, approaches crime analyses based on "new" knowledge, which includes digital intelligence, data and information processing monitoring and demographic analysis to map crime patterns and methods. Several other similar Knowledge based systems or KBSs have been implemented in recent years in this field. These include the Integrated Ballistics Identification System (IBIS), Automated Palm and Fingerprint Identification System (AFIS), COPLINK Connect and POLNET (Donalds et. al, 2019). All the above-mentioned frameworks are, however, primarily focused on regular crimes, with very limited coverage of cybercrime in its entirety. With the advent of the Internet, connectivity, and ease of use of personal electronic devices, criminals are now committing traditional and new crimes using technology. According to Wall and April (2005), the Internet has enabled the rise of three aspects of criminal opportunity: traditional crimes such as

fraud, money laundering, pyramid schemes, stalking and sexually motivated crimes using technology, new criminal activities such as Hacktivism, identity theft, 419 scams as well as opportunities for new types of digital crimes such as e-auction scams, spam, DoS and DDoS attacks, hate speech and intellectual property infringement and piracy. Hence the new push for KBS is to develop comprehensive frameworks that include the occurrence and trends of cybercrime. The most popular form of cybercrime representation to date has been using ontological models for cybercrime classification. The Crime emergency Event Model (CE2M) ontology was developed to implement semantic level integration of all crime emergency service resources. A four-dimensional approach was adopted by Hansman and Hunt (2005) to represent a computer and network attack classification taxonomy. Kjaerland (2005) used factors such as source sector, impact, target, and method of operation as a basis for classifying cybercrimes. In relatively recent years, the cybercrime taxonomies have been targeted at categorizing their specific properties rather than the actual cybercrimes themselves. elements such as victim, attacker, objective, tool and tactic, impact, result, relationship, target, and offense are typically used to classify cybercrimes in an ontological domain. Although current cybercrime classification taxonomies assert a comprehensive categorization of cybercrime characteristics, one aspect of these crimes have been heavily neglected.

We believe that to provide a complete, holistic and a user centric perspective for current cybercrime trend, the inclusion of a complete victimology profile of the user of the digital world is of utmost importance. Understanding the victimology behind such incidents is effective in developing a complete picture of the cybercrime event. Hence, a comprehensive analysis of cybercrime victimology is required to completely understand cybercrimes from the perspective of the user as active vulnerability points. To assess a complete picture of cybercrimes as crimes mediated by computer networks rather than crimes against machines, it is necessary to analyse these crimes from different viewpoints. The focus of crime perspective has always typically been from the point of view of the offender – the malicious actor or cybercriminal. However, a new approach in the current millennium cybercrime definitions is moving beyond the machines, the attacker, and focusing on the target humans, especially the victims. A new victimization typology of cybercrime has been proposed by Jaishankar (2012, 2013) that identifies four targets of cyber victimization – Cyber victimization of Governments, Cyber Victimization of Corporations, Cyber Victimization of Individuals or Interpersonal Cyber Victimization and Victimless Crimes. Riding on this new approach and adopting the current cybercrime characteristic based ontological representation, this paper proposes a novel outlook to cybercrime victimology-based ontology that enhances present cybercrime taxonomies to provide an in-depth look through the perspective of the victim.

The rest of the paper is organized as follows: Section 2 provides a brief glimpse into various knowledge-based taxonomies of cybercrimes and the popular ontological representations of cybercrime characteristics. In Section 3, the paper delves into the OWL ontological representation which inspires our work along with a discussion of the victim typology classification. Section 4 describes how our model extends the previous representation to include a comprehensive victimology perspective and demonstrates the resultant ontological model for cybercrime victimology. Section 5 delves into an example application of our Victimology categories in real-world cybercrime scenarios and discusses the limitations and improvements needed. We conclude our findings in Section 6 with a discussion of our proposed future work.

2. Related Research

Previous research has investigated several models of cybercrime classification. The Hansman and Hunt (2005) classification taxonomy included four dimensions and a specific vulnerability dimension and attack target categorization. Although the elements proposed in this model are effective in classifying traditional cybercrimes, it fails to include blended attacks. Choo (2011) defines blended attacks as a technique which utilizes both semantic (exploiting human vulnerabilities) and syntactic (exploiting technical vulnerabilities) attack methods against the victim. Inclusion of this category of blended attacks contribute to a more comprehensive cybercrime ontology.

The four facetted cybercrime classification scheme was also utilized by Kjaerland (2005) to analyse attacks against government and commercial sectors. It was successful in identifying new characteristics for cybercrime classification, although it was limited in the inclusion of very few characteristics of cybercrime. Thus, this taxonomy by itself fails to provide the comprehensive outlook we are striving for.

Simmons et. al. (2009) moved away from the four-dimensional approach to a five characteristic based approach of cybercrime taxonomy. Coined AVOIDIT, this taxonomy uses Attack Vector, Operational Impact, Defence, Information Impact, and Target as the five dimensions. A noticeable improvement on previous methods was the inclusion of blended cyber-attacks as well as the analysis of the relationships between targets, hosts, access

paths and impacts. Although this taxonomy is a significant improvement over its predecessors, it still lacks a complete approach towards cybercrime, as it does not consider the attack classification, motive or objective of the attacker nor identifies the target victim of these attacks. Both these categories are important factors to be considered in providing a complete picture of cybercrime patterns.

In the domain of ontological framework development of cyber-attacks, van Herdeen, Irwin, Burke, ad Leenen (2012) proposed a taxonomy to represent network attacks. The resultant ontological framework utilized classes of attack scenario, actor, actor location, aggressor attack goal, motivation, scope size, scope, target, vulnerability, asset, sabotage, effect, phase, attack mechanism and automation level. The framework based on computer network attacks proved to be beneficial in identifying additional cybercrime characteristics such as the vulnerability and aggressor classes. However, the ontology suffers from the limitation that it does not encompass all types of cybercrimes. It fails to include blended attacks as well as victim specific targeted attacks, necessary for the comprehensive picture.

The categories and subjects' approach were adopted by Applegate and Stavrou (2013) to describe a cyber conflict taxonomy. Subjects were identified as entities or events which represent real world cyber conflict events and its associated participants. Categories included two subcategories of actions and. Actors, which were further subdivided into further subcategories. This taxonomy served as an extension on the Simmons et al. version and was successful in refining more specific categories for cybercrime classification, along with the relationships among the elements. However, this taxonomy still suffers from its shortcomings in terms of the thoroughness of insights into a cybercrime incident. It fails to include attacker objective, the specific attack or offense and the resultant victim or complainant impacted, all of which are significant in building the complete cybercrime outlook.

The victim, attacker, objective, tool & tactic, impact, result, relationship, target, and offense characteristics approach was further adopted by Donalds and Osei-Bryson (2014) in their taxonomy focused on cybercrime properties rather than the nature of the crime. This approach enables identifying links between taxonomic characteristics and successfully integrates and extends previous classification schemes. Other relevant characteristics of cybercrime, namely attack event vulnerability and impact types were not considered as part of this taxonomy and hence were limited in these aspects.

An integrated view of cybercrimes was proposed by Barn and Barn (2016) and implemented in Protégé OWL (Web Ontology Language), which included a comprehensive classification of cybercrime characteristics. A variation from previous taxonomies was the inclusion of the "viewpoint of external observers", which was significantly different from a law enforcement and policing based on CSIRT analyses. Moreover, the ontology combined traditional and cybercrime characteristics, which many researchers felt were easily captured using a technology subclass and not necessarily needed as a separate taxonomy element. Finally other relevant cybercrime concepts such as vulnerability, operational and informational impacts were not incorporated.

In terms of the role and impact of the user or human in cybercrimes, the research has been somewhat limited. Islam et. al. (2019) believes that the context of users and their relationships with the devices are mainly focused on the awareness and incentivization content. It fails to capture the ever-evolving user preferences and innovative attackers in a constantly changing cybersecurity and cybercrime landscape. The authors thus proposed a holistic socio-technical framework combining relevant concepts such as opportunity management, behavioural and business models. This framework aims to reduce human enabled risks in the cybercrime scenario by considering human roles of offenders, victims, preventers, and promoters. Although the framework considers the different user roles in cybercrimes, the focus is still on the attacker perspective and opportunity to reduce human related vulnerabilities of a system.

Philips et. al. (2022) in their research highlighted the importance of a comprehensive and standardized terminology and categorization of cybercrime elements. The push is towards building a comprehensive framework to conceptualize cybercrime after evaluating the existing definitions, typologies, and taxonomies prevalent in various law enforcement jurisdictions. However, the authors recognize that the resultant classification approaches come with their own weaknesses, specifically the failure to capture a holistic picture of the cybercrime. The authors believed that using a human perspective such as the motivations and intentions of the attacker may yield additional knowledge points in the framework representation.

The need for analysis of the human component of cybercrimes has led to increased attention to the cybercriminology division. Subramaniam et. al. (2022) observed several cybercrime frameworks and the lack thereof in different countries such as Malaysia and Nigeria and identified the need for a cyber victim intervention

framework along with a cyber security framework. Recognition of cyber victims and victim studies were deemed to be significant aspects in the goal towards holistically protecting our cyberspace. Thus, our paper focuses on the importance of understanding cybervictimology and proposes a comprehensive ontological framework which includes the role and impact of cyber victims.

With the above discussion, it is evident that prior attempts at cybercrime classification have largely focussed on the different concepts of classification. These works have also used varying terminologies interchangeably, leading to a unified and standard approach for such taxonomic frameworks. Moreover, these non-adaptable frameworks are unable to account for ever changing dynamic and information for cybercrime incidents. This research also fails to demonstrate the complex relationships and interactions between framework elements as well as lacking the ability of the analysis of cybercrimes from varied perspectives.

Our proposed research aims to address all these shortcomings in our modified victimology-based taxonomy approach.

3. Background

This section explores the relevant background for the proposed research. We delve into the origin of the term Cybervictimology and its associated connotations. Further, we observe a typology for victims as published in previous research. This typology includes victim dimensions and their categorizations. Although the typology has been proposed for analysing victim profiles for traditional crimes, we draw a parallel of these classifications as applicable in our proposed framework of cybercrime specific victimization.

3.1 Cybervictimology

The terms cybercriminology and cybervictimology were first coined by K. Jaishankar in 2007 and 2015 respectively. Cybercriminology is defined to be a sub area of criminology, especially crimes involving the internet. The resultant victims of these crimes are considered cyber victims, and hence their study is known as cybervictimology. Cybercrimes are known by various terms such as digital crime, internet crime, virtual crime, e-crime or netcrime, they all represent the use of internet and communication networks to commit these crimes. However, compared to traditional crimes, victimization in cybercrimes is especially significant and distinct since, the cyber criminals do not necessarily need to be physically present or even close to the victims. An attacker can victimize any person anywhere in the world, just by using his devices and the internet. Also, the issue of timeline of the crime gets blurred as cybercrimes can automated and hence can occur over a given time range. Moreover, one crime can result in victimization of multiple people at a time with the same effort. Thus, the concept of victimization in the cybercrime context becomes important due to the nature of automation and the ability of offenders to impact rapidly and precisely many individuals at a time.

The Convention on Cyber Crime in the Council for Europe (2011) was one of the first entities who passed a legislation to consider cybercrime definitions from the victims (humans' perspective rather than concentrating solely on the machines). The Convention identified offenses against children, attacks targeting human emotions, and the ban on improper words in cyberspace, instigating threats towards national security, racial hatred, and terrorism. Moreover, the perspective in identifying cybercrimes was shifted from only hacking or e-commerce attacks to include emotional attacks on users, and to evaluate attacks from the victim's perspective. Wall (2008) first included the component of "harm" and "online insecurity and risk" in cybercrime definitions. This enabled the evaluation of impact of crimes like stalking, harassing, and bullying against human emotions. Cyber victimization, whenever recognized has also have had only a singular focus from the individual's perspective and not from a mass or system perspective including governments, corporations, and society. Hence, considering these aspects is crucial for a complete definition of cyber victimization. Finally, the response of the victim and its effects in the overall cyberspace, which enables investigation of causes of re-victimization.

Haider and Jaishankar (2012) have since considered cybercrime from the victimization perspective and defined the terminology as follows:

> *"Offences that are committed against individuals or groups of individuals with a criminal motive to intentionally harm the reputation of the victim or cause physical or mental harm to the victim directly or indirectly, using modern telecommunication networks such as Internet (Chat rooms, emails, notice boards and groups) and mobile phones (SMS/MMS)."*

The above definition encompasses all the following cybercrimes and associated abetment, towards individuals:

Hacking, Morphing, Spoofing, Computer source tampering, Obscene publication, Trojan attacks, Phishing, Cyber stalking, Cyber pornography, Cyber defamation, Cyber bullying, E-mail harassment, Cyber blackmailing, Cyber threats, Cyber murder, Cyber terrorism

Recognizing the wide array of cybercrime victimization aspects, our work seeks to demonstrate an easy-to-use model to explore the relevant elements in a typical cybercrime scenario and the association between the elements as viewed from the victim perspective.

3.2 Victim Typology

Landau and Freeman-Longo (1990) published research outlining a comprehensive classification of victim dimensions along with their different categorizations. We now take an in-depth look at the typology of a victim typically in the context of traditional crimes (Figures 1a and 1b). Our proposed research analyses elements from these typologies and implements the applicable ones in the final ontological model.

A Conceptual Framework for Victimology: The Dimensions and their Categories

Dimensions	1	2	3	4	5	6	7	8
A. *Source of Victimization*	Individual	Group	Community	State	Corporate	Technological Environment	Interface Environment	Natural Environment
B. *Legal Framework*	Criminal Law	Civil Law	Transnational, International Law	No Legal Framework				
C. *Intentionality of Perpetrator*	Intentional	Recklessness	Negligence	Accident	Not Applicable			
D. *Identification of the Victim*	Individual	Social Group	Corporate	Nation(s)	Mankind in General			
E. *Victim Vulnerability*	Age	Sex	Biophysical Charac-teristics	Psychological Charac-teristics	Social Charac-teristics			
F. *Victim's Perception of Victimization*	Accurate self-perception of victim status	Inaccurate self-perception of victim status	Non-perception of victim status	Ignorance of the victimization				

Figure 1a: Dimensions and categorizations in a Victimology framework (Landau and Freeman-Longo, 1990)

G. *Other's Perception of Victimization*	Accurate perception of victim status	Inaccurate perception of victim status	Non-perception of victim status	Ignorance of the victimization			
H. *Type of Victimization*	Physical Harm/ Damage	Sexual Abuse	Economic Damage	Psychological Damage	Damage to Reputation	Infringement of Civil/ Human Rights	
I. *Severity of Victimization/ Harm*	None	Mild	Moderate	Severe	Extreme	Maximal	
J. *Victim-Offender Relationship*	Family	Acquaintance	Professional	Stranger	Impersonal	Not Applicable	
K. *Victim's Contribution to the Event*	None	Minimal	Moderate	High	Maximal		

Figure 1b: Dimensions and categorizations in a Victimology framework (Landau and Freeman-Longo, 1990) (continued)

3.3 Knowledge-Based Cybercrime Ontology

Donalds and Osei-Bryson (2012) presented a comprehensive cybercrime ontology based on the different elements present in a typical cybercrime scenario. The model identified the primary components for classification of cybercrime incidents the relevant relationships between them. The model was proposed to be extensible so that additional cybercrime characteristics and associated relationships can be added later an as-needed basis. Table 1 demonstrates the components of this cybercrime ontological model as proposed by the authors. The first column outlines all the elements discussed in the said ontology. The second column provides the definitions provided by the authors for these elements. Columns three and four demonstrate which of the elements have been utilized in our new extended cybervictimology model. As observed, for the purposes of this paper, we focus only on the 'Victim' element from the original model and discuss the typological classifications adopted and implemented in our ontology.

Table 1: Cybercrime classification ontological concepts

Element Name (Old Model)	Definition (Old Model)	Use in New Model (No Change)	Use in New Model (Changed
Attack_Event	*"Actual real-world cybercrime or nefarious cyber action committed by an Attacker"*	✓	
Vulnerability	*"Actual real world cybercrime or nefarious cyber action committed by an Attacker"*	✓	
Tool_and_Technique	*"Tool and/or technique used by an Attacker to execute a cyber action"*	✓	
Objective	*"The main purpose, motive or end goal of an Attacker for committing a cyber event"*	✓	
Offence	*"An illicit cyber action or event"*	✓	
Location	*"Country and/or specific address of an individual/group Attacker, the Victim and/or Target that experiences the event"*	✓	
Complainant	*"One who reports a cyber event, making the authorities aware of the illicit activity"*	✓	
Victim	*"An entity that is affected in some way by a cyber event"*		✓
Target	*"An entity that a cyber event is specifically directed at"*	✓	
Impact	*"The direct effect caused by a cyber event on a Victim"*	✓	
Attacker	*"Criminals trying to exploit resources for personal, governmental and financial goals. "*	✓	

4. Our Approach

In the subsequent sub-sections, we discuss the proposed approach of building an ontology based on the cybervictimology classifications adopted.

4.1 Victimology-Based Cybercrime Approach

As mentioned previously, we adopt the CCO (Cybercrime Ontological) model, and its associated elements as defined in the ontological framework. Figure 2 shows the complete ontological representation as discussed in Donalds and Osei-Bryson (2012). The 'Victim' component (boxed in red colour) is the only one we have expanded as part of our current research. Since the purpose of our approach is to gather a different victim-based perspective on cybercrimes and their impact, the comprehensive expansion of this ontological element is the prime point of focus.

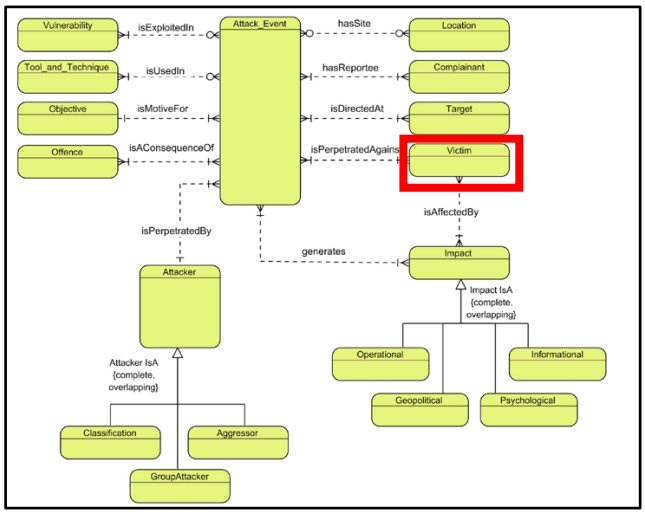


Figure 2: Cybercrime Ontological Model (Donalds and Osei-Bryson, 2012)

4.2 Cybercrime Victimology Ontological Model

To understand and implement effective policing for cybercrimes, there is a need for knowledge-based systems and applications. However, such methods specifically in cybercrime is largely lacking. Even with a few such frameworks in use, a comprehensive coverage of all elements of cybercrime and evaluation of different perspectives in the same are unexplored. In this research, we adopt an existing cybercrime knowledge-based

taxonomy and further enhance it by adding elements of cyber victimology. This enables the introduction of a victim-based perspective in cybercrime. We develop an ontological model of our modified victimology-based framework. An ontology is an efficient mechanism to capture knowledge in a particular domain and helps in conceptualizations of the domain elements into a human-understandable, but also a machine-readable format using objects, relationships, attributes, and axioms.

Our proposed cybercrime victimology framework considers all the above-mentioned victimization categorizations in its original for4mat. However, the only place it differs from the original model is the categorization of Victims. We utilize the typology of victims outlined in Section 3.2 and expand the Victim attribute in the ontological model with its own categorizations and definitions as required. Although Section 3.2 has 11 categorizations (A-K) in the Victim typology, for our purposes of cybercrime victimization representation, we adopt a select set of these categories. We did not include the dimensions of Intentionality of the perpetrator, Victim's perception of Victimization, and Others perception of Victimization, as these dimensions are more applicable to traditional crimes and victims rather than cybercrime. We assume that the cybercrime perpetrator has the intention of carrying out his actions and makes conscious choices to exploit vulnerabilities. We also remove the possibility of ambiguity in victim perception and for the purposes of our model assume a guaranteed impact of cybercrime events on the victims involved directly and indirectly. We represent our ontological model in Protégé OWL capturing the victimology aspects of cybercrimes as the expansion. The resultant ontological model for cybercrime victimology is provided in Figure 5. This model can be added on to the overall knowledge-based system model for cybercrimes discussed previously.

The 8 Cyber victimology typology dimensions and their associated categories, as adopted in the proposed model are as follows.

1. *Source of Victimization* (Individual, Group, Community, State, Corporate, Technological Environment, Interface Environment, Natural Environment)
2. *Legal Framework* (Criminal Law, Civil Law, Translational, International Law, No Legal Framework)
3. *Identification of the Victim* (Individual, Social Group, Corporate, Nation(s), Mankind in general)
4. *Victim Vulnerability* (Age, Sex, Biophysical characteristics, psychological characteristics, social characteristics)
5. *Type of Victimization* (Physical Harm/Damage, Sexual Abuse, Economic Damage, Psychological Damage, Damage to Reputation, Infringement of Civil/Human Rights)
6. *Severity of Victimization/Harm* (None, Mild, Moderate, Severe, Extreme, Maximal)
7. *Victim-Offender Relationship* (Family, Acquaintance, Professional, Stranger, Impersonal, Not Applicable)
8. *Victim's Contribution to the Event* (None, Minimal, Moderate, High, Maximal)

Figure 3 shows the victimology typology dimensions that were adopted from Donalds and Osei-Bryson, 2012. The dimensions are represented as object classes and their relevant categories are represented as subclasses respectively.

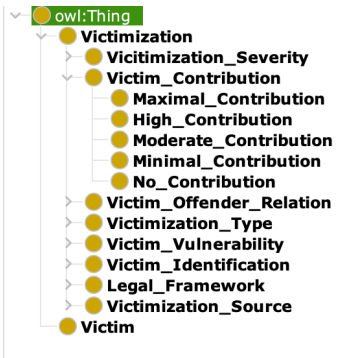


Figure 3: Victimology dimensions and categorizations as ontological class objects

Figure 4 demonstrates the ontological representation of the relationships between the object classes. We define 9 new object properties to represent the relationships between each of the victimology dimensions and their respective categories.

1. Victimization ------←--------isRelatedAs---------←------ Victim_Offender-Relation
2. Victimization ------←--------hasInvolvement---------←------ Victim_Contribution
3. Victimization ------←--------causesResults---------←------ Victimization_Type
4. Victimization ------←--------hasIntensity---------←------ Victimization_Severity
5. Victimization ------←--------isIdentifiedAs---------←------ Victim_Identification
6. Victimization ------←--------isGovernedBy---------←------ Legal_Framework
7. Victimization ------←--------isInitiatedBy---------←------ Victimization_Source
8. Victimization ------←--------isCausedBy---------←------ Victim_Vulnerability
9. Victimization ------←--------undergoes---------←------ Victim

These object properties represent the following relationships between the victimology typology dimensions and their categorizations. The final ontological representation of the Cybercrime Victimology representation is illustrated in Figure 7 (provided at the end of the paper for better visibility).

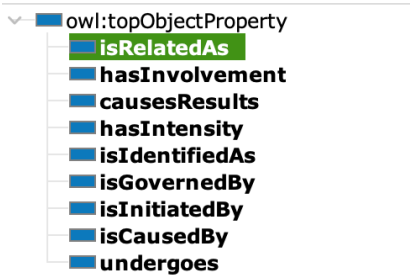


Figure 4: Relationship between Victimology dimensions and respective categorizations as object properties

5. Classification of two Real World Cyber Attacks

In this section, we outline two real-world cybercrime cases as documented in the Sherloc UNODC database (to https://sherloc.unodc.org/). We highlight the relevant portions of the cases below for the sake of brevity. For the detailed case reports please go to https://sherloc.unodc.org/. We illustrate three more cases from the case law database in the appendix section of the paper. We focus on the victimization aspect only for these examples. The other attack aspects are covered in a similar manner to the Donalds et.al 2012 classification model. The following victimization dimensions and their corresponding categorizations were identified as follows:

5.1 Case DOMx001 (Identity Offence)

DOMx001 **Segundo Tribunal Colegiado De La Cámara Penal Del Juzgado De Primera Instancia Del Distrito Nacional, Sentencia penal núm. 249-04-2021-SSEN-00225**

Dominican Republic

Verdict Date: 2021-11-04
Sentence Date: 2021-11-04

In 2018, the two defendants, Eric Antonio Morel Garcia and Henry William Cruz, along with others (Rafael Leónidas Sepúlveday and Winifer Hernández) defrauded the victim by fraudulently obtaining access codes to her bank accounts, which allowed them to conduct multiple electronic transfers of fraudulent funds in the total amount of DOP 2,336,000.00. The victim was contacted by a person who self-identified as Doña Carmen, pretending to be the person who will manage her Cibao Savings and Loans Association accounts. "Doña Carmen" informed the victim that she should go to the bank branch and request a code card to activate Internet Banking. The victim was subsequently contacted multiple times by "Doña Carmen" requesting the card code access, under the pretext that the person was adjusting her bill. When the victim received credit card transactions she contacted "Doña Carmen" who told her it was a platform problem.

Victimization Analysis	
Source of Victimization	Group
Legal Framework	Criminal Law
Identification of the Victim	Individual
Victim Vulnerability	Inadequate data
Type of Victimization	Economic Damage
Severity of Victimization/Harm	Moderate
Victim-Offender Relationship	Stranger
Victim's Contribution to the Event	Moderate

Figure 5: Case DOMx001 summary (https://sherloc.unodc.org/)and categorization

5.2 Case USAx209 (Forgery)

USAx209 **United States of America v. Khan, No. 17-4301 (4th Cir. Apr. 4, 2018)**

United States of America

Verdict Date: 2020-04-04
Sentence Date: 2018-04-04

This case involves several offences, including conspiracy, smuggling cultural property into the U.S., obstruction of justice, and citizenship and naturalization fraud, committed by the defendant Ijaz Khan with the assistance of several co-conspirators. In 2002, the co-conspirator Vera Lautt travelled to Pakistan to meet Ijaz Khan, whom she had met online, and to sign marriage documents. Subsequently, they both submitted fraudulent documents to the U.S. Department of State (DOS) and U.S. Citizenship and Immigration Services (USCIS), facilitating the immigration and later naturalization of Ijaz Khan. The defendant Ijaz Khan was, however, still married to a Pakistani woman, with whom he had and continued to have children, at the time of and after signing the marriage documents. According to the indictment, Ijaz Khan used his citizenship to enable the immigration and later naturalization of his four oldest sons. He divorced Lautt before the children had arrived in the United States. He further filed or assisted with petitions on behalf of his brother, Ibar Khan, his Pakistani wife, his mother, and another two children.

Victimization Analysis	
Source of Victimization	Group
Legal Framework	Civil Law, International Law
Identification of the Victim	Social Group, Nation
Victim Vulnerability	Sex, Psychological characteristics, Social characteristics
Type of Victimization	Psychological Damage, Infringement of Civil/Human Rights
Severity of Victimization/Harm	Maximal
Victim-Offender Relationship	Family
Victim's Contribution to the Event	Moderate

Figure 6: Case USAx209 summary (https://sherloc.unodc.org/)and categorization

5.3 Observations and Inferences

From an initial analysis of the above-mentioned real-world cybercrime cases and similar others from the database, we can highlight the following observations:

- The identity and details of the victim(s) are often hidden in traditional crimes and this practice has been present in cybercrimes as well. Hence, sufficient information on Victim characteristics such as Age, Sex, Social, Economic background, etc., has been largely missing from the cybercrime database case descriptions. Hence this categorization has not yielded the necessary information and the subsequent pattern expected from cybercrimes. Access to more detailed case descriptions from the law enforcement and legal case files are needed to make an accurate observation in this case.
- As with traditional crimes, the categorizations of cybercrime types into identity thefts, fraud, forgery, etc. are not mutually exclusive. Each of the cybercrime cases have elements of different categorizations that blur the lines to make effective demarcations of crime types difficult. For example, the case of DOMx001 has potential for charges under fraud as well as Identity offenses. Hence, either an overlap criterion for documenting these cases in an ontology should be studied and expressed appropriately in a framework.
- The details of the geographical extents of the victims are also missing in most cybercrime cases in the open-source databases. Like victim identities, their geographical locations can be useful in deciding applicability of local and international laws, relationship of victim and perpetrators, Victimization severity and Victim involvement. Detailed knowledge of these victim dimensions will provide a more complete picture of the crime from the victims' perspectives, that will in turn enable possible patterns between the victims' location, ease of access for the crime, prior relationship to the perpetrator, attempted social engineering attacks if any as well as legal precedents of similar cases from other countries if needed.
- The Victim dimensions and their corresponding categorizations have been effectively defined by Landau and Freeman-Longo (1990). However, more refined, and targeted definitions of these categories need to be established to fit the varying nature of cybercrime. As we observe, cybercrime victimization can span countries, different socio-economic groups and fall under different legal jurisdictions. Cybercrimes may also be encountered as a means or part of a traditional crime scenario and need a different approach of legal applicability and indictment. Hence, refining the provided definitions of victimizations to suitably fit cybercrime is justifiably needed. For example, Case USAx209 victims can be identified as both social group (the defendant's original family in Pakistan) as well as the nation state of USA. Hence the justification of applying one or the other as part of the 'Identification of Victim' dimension are required for the ontological representation.

6. Conclusion and Future Work

In this paper, we outlined a taxonomic framework of cybercrime elements from the perspective of the victim. Including the victimology viewpoint in cybercrime analysis has received significant importance in recent times and a complete understanding of victimology patterns and details provide a holistic picture of a cybercrime classification system. We represented our model in Protégé OWL which provides an easy to visualize and explorable structure for different cybercrime elements. We followed a classic victimology typology and selected specific categories from it as appropriate for a cybercrime environment. We implemented a preliminary analysis of our model with real-world cybercrime descriptions. As part of our future work, we will attempt to generate an ontological representation of the incidents in a case study format. And observe and analyse victimology patterns to better understand and perfect cybercrime classification from the victimology perspective. Further work in this area will focus on developing SWRL rules to enable querying specific cybercrime details from the ontological model relevant to the case studies represented.

References

Applegate, Scott D., and Angelos Stavrou. "Towards a cyber conflict taxonomy." *2013 5th International Conference on Cyber Conflict (CYCON 2013)*. IEEE, 2013.

Barn, Ravinder, and Balbir Barn. "An ontological representation of a taxonomy for cybercrime." (2016).

Blythe, John M., and Lynne Coventry. "Costly but effective: Comparing the factors that influence employee anti-malware behaviours." Computers in Human Behavior 87 (2018): 87-97.

Choo, Kim-Kwang Raymond. "The cyber threat landscape: Challenges and future research directions." Computers & security 30.8 (2011): 719-731.

David, Wall, and Pattavina April. "The Internet as a conduit for criminal activity." Information technology and the criminal justice system. Sage, 2005. 77-98.

Wall*, David S. "Cybercrime, media and insecurity: The shaping of public perceptions of cybercrime." International Review of Law, Computers & Technology 22.1-2 (2008): 45-63.

Donalds, Charlette, and Kweku-Muata Osei-Bryson. "A cybercrime taxonomy: case of the Jamaican jurisdiction." (2014).

Donalds, C. M., and K. M. Osei-Bryson. "The construction of A domain ontology for criminal investigation: The case of the Jamaican constabulary force." SIG ICT in Global Development, 5th Annual Pre-ICIS Workshop, Orlando, FL. 2012.

Donalds, Charlette, and Kweku-Muata Osei-Bryson. "Toward a cybercrime classification ontology: A knowledge-based approach." Computers in Human Behavior 92 (2019): 403-418.

Guarino, Nicola, and Pierdaniele Giaretta. "Ontologies and knowledge bases." *Towards very large knowledge bases*(1995): 1-2.

Hansman, Simon, and Ray Hunt. "A taxonomy of network and computer attacks." *Computers & Security* 24.1 (2005): 31-43.

Van Heerden, Renier P., et al. "A computer network attack taxonomy and ontology." International Journal of Cyber Warfare and Terrorism (IJCWT) 2.3 (2012): 12-25.

Islam, Tasmina, et al. "A socio-technical and co-evolutionary framework for reducing human-related risks in cyber security and cybercrime ecosystems." *International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*. Singapore: Springer Singapore, 2019.

Jaishankar, K. "Cyber victimology: a new sub-discipline of the twenty-first century victimology." *An International Perspective on Contemporary Developments in Victimology: A Festschrift in Honor of Marc Groenhuijsen* (2020): 3-19.

Keyser, Mike. "The council of Europe convention on cybercrime." Computer Crime. Routledge, 2017. 131-170.

Kjaerland, Maria. "A taxonomy and comparison of computer security incidents from the commercial and government sectors." Computers & Security 25.7 (2006): 522-538.

Landau, Simha F., and Robert E. Freeman-Longo. "Classifying victims: A proposed multidimensional victimological typology." *International review of victimology* 1.3 (1990): 267-286.

Phillips, Kirsty, et al. "Conceptualizing cybercrime: Definitions, typologies and taxonomies." *Forensic sciences* 2.2 (2022): 379-398.

https://sherloc.unodc.org/

Simmons, Chris, et al. "AVOIDIT: A cyber-attack taxonomy." *University of Memphis, Technical Report CS-09-003* (2009).

Subramaniam, S., Ratha Krishnan, S. & Sathakathulla, S.A. 2022. Cyber Victim Intervention Framework: An Introduction. Selangor Women Conference 2022. Selangor Journal.

Tsakalidis, George, et al. "A cybercrime incident architecture with adaptive response policy." *Computers & Security* 83 (2019): 22-37.

Appendix 1

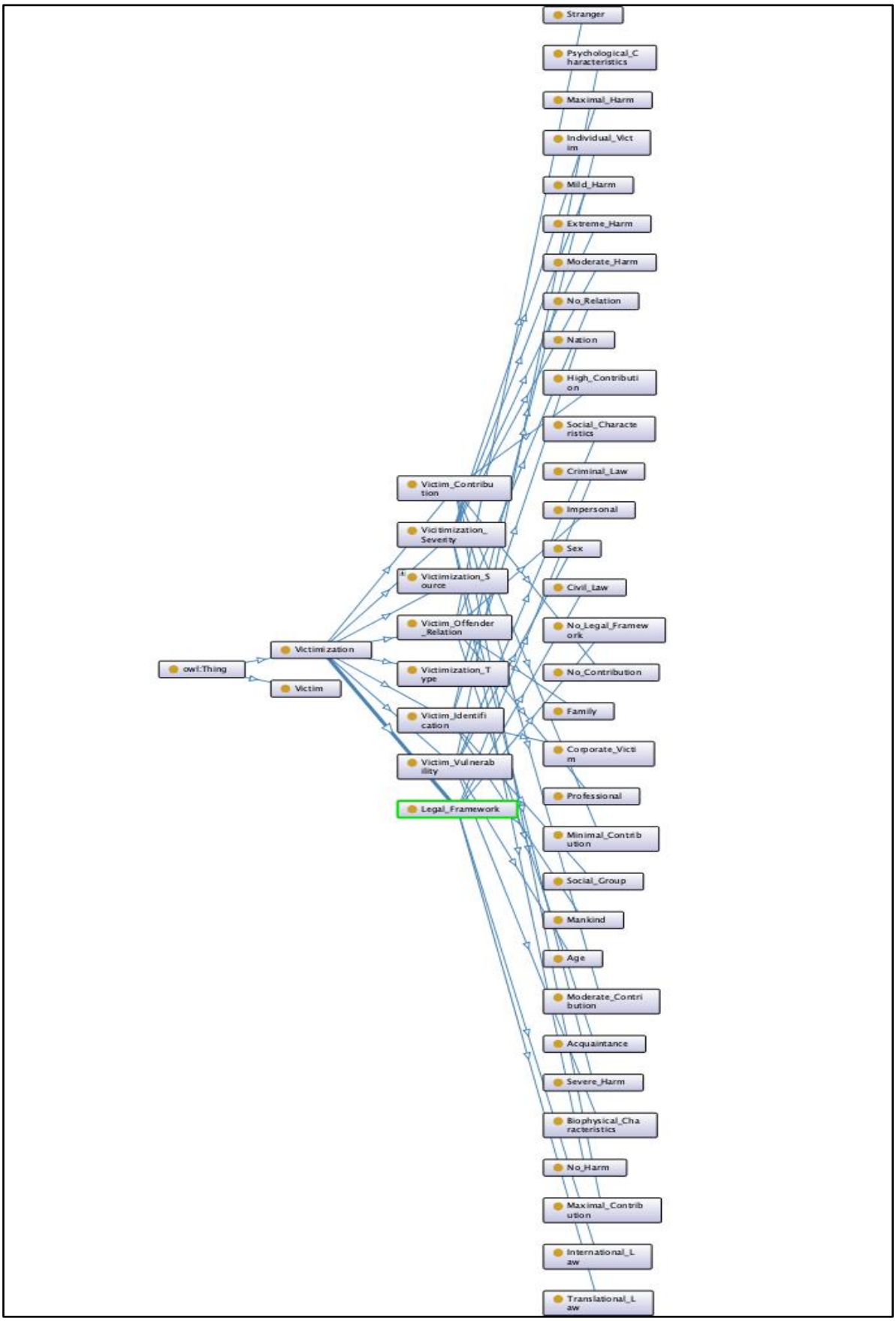


Figure 7: OWL ontological representation of victimology dimensions and categorizations

Case FRAx030 (Forgery)

FRAx030 **TGI Paris, 13e ch. corr., jugement du 20 novembre 2018**

France

Verdict Date: 2018-11-20
Sentence Date: 2018-11-20

From 2010 to 2014, the defendant Mr Z. conducted a criminal enterprise to fraudulently purchase goods on commercial websites. To this end, Mr Z. used stolen credit card data found on carding forums by himself, Mr P. (the technical advisor of the group) and Mr N. (a member of the group in charge of finding credit card data), as well as credit card data that Mr L. had stolen from his former employer. Mr P. and Mr Z. would then hack into customer accounts on commercial websites and modify their contact information so that the actual customer would not receive any notification of purchase and/or delivery. Mr Z. and N. would then buy goods on commercial websites and send them to parcel collection points. Mr Z and Mr X forged fake IDs and used mules to receive the packages at the collection points. Several people were used as mules: Messrs Y, M, O, Q, V, T, and R. The mules would receive the packages, keep some packages as payment and send the others to Mr Z. so that he could sell them on retailer websites. Several people involved in this criminal organization later started using the same techniques to buy goods on their own behalf.

Victimization Analysis	
Source of Victimization	Group
Legal Framework	Criminal Law
Identification of the Victim	Social Group, Corporate
Victim Vulnerability	Social Characteristics
Type of Victimization	Economic Damage
Severity of Victimization/Harm	Severe
Victim-Offender Relationship	Impersonal
Victim's Contribution to the Event	Minimal

Case ARGx016 (Fraud)

ARGx016 **Juzgado En Lo Correccional Nº 1 - San Isidro, Case No. SI-3862-2021**

Argentina

Verdict Date: 2022-03-22
Sentence Date: 2022-03-22

A criminal group, including the defendant XXX, committed fraud by masquerading as a Banco Galicia employee. The group sent an Instagram message to the victim informing her that if she wanted to receive advice from the financial institution, she should provide her cell phone and her area code. She subsequently received calls from two numbers from the Province of Córdoba via WhatsApp from someone pretending to be an employee of the Bank Galicia (this person was still unidentified at the time of this case). The person who called the victim tricked her into providing her banking details and her security token. The victim's information was then used to obtain a loan in the amount of ARS 189,448.00, which was later transferred together with the additional money contained in the victim's account, making it a total of ARS 229,000.00, to an account in the name of the defendant. The money was then transferred to another account of the defendant and an account of XXX. They in turn then transferred the money to others (YYY, ZZZ and OOO).

Victimization Analysis	
Source of Victimization	Group
Legal Framework	Criminal Law
Identification of the Victim	Individual
Victim Vulnerability	Psychological characteristics
Type of Victimization	Economic Damage
Severity of Victimization/Harm	Maximal
Victim-Offender Relationship	Stranger
Victim's Contribution to the Event	High

Case USAx217 (Identity Offence)

USAx217 **United States of America v. Wyatt, No. 4:17-cr-00522-RLW-SPM (E.D. Mo. Sept. 21, 2020)**

United States of America

Verdict Date: 2020-09-21
Sentence Date: 2020-09-21

This case involves a member of the cyber organized criminal group "The Dark Overlord" (TDO), an organization that was known for their ransom and extortion scams. Beginning in 2016, the defendant was involved in a conspiracy that encompassed hacking several organizations in health, entertainment, finance, commercial, real estate, and transportation, stealing personal information from the systems they hacked, including for instance medical records, and then seeking ransom from the targets. Companies would usually have to pay a ransom of between USED 75,000 and 350,000 in bitcoin. The defendant's role in the conspiracy was creating, validating, and maintaining communication, payment, and virtual private network accounts that were, inter alia, used to send threatening and extortionate messages to the victimized companies within the Eastern District of Missouri, where the defendant was ultimately tried.

Victimization Analysis	
Source of Victimization	Group
Legal Framework	Criminal Law
Identification of the Victim	Corporate
Victim Vulnerability	Inadequate data
Type of Victimization	Economic Damage, Damage to Reputation
Severity of Victimization/Harm	Severe
Victim-Offender Relationship	Professional
Victim's Contribution to the Event	Maximal

Chain of Custody and Evidence Integrity Verification Using Blockchain Technology

Adir Miller and Avinash Singh
University of Pretoria, South Africa

u20692286@tuks.co.za
asingh@cs.up.ac.za

Abstract: The validity and integrity of digital evidence and the chain of custody are crucial to all digital forensic investigations. All new evidence and access logs of the original evidence should be logged in a document called the 'chain of custody'. This document shows the timeline of any piece of evidence from the time it was recorded until the end of the investigation. In a traditional digital investigation, trusted parties, such as an investigator, are allowed access to the digital evidence and follow a strict process when dealing with data. These trusted parties have the capability to alter the data making the evidence inadmissible in a court of law. Alternatively, these trusted parties may also alter the data accidentally or with malicious intent, due to a lack of transparency and non-repudiation. Blockchain technology can solve this issue, however, existing research shows that adopting blockchain does not provide adequate transparent access control mechanisms. Consequently, this makes blockchain difficult to adopt due to the one-to-one mapping and the inability to easily validate the chain of custody and evidence admissibility. Current methodologies rely on an external off-chain access control mechanism, which, regrettably, remains susceptible to potential breaches that could compromise its integrity and validity. This paper proposes an enhanced model to provide access control through smart contracts, ensuring immutability, flexibility, transparency, and non-repudiation of both the access control mechanisms and the digital evidence itself. This is achieved by moving the access control mechanism to the blockchain. This tracks any changes made through the access control mechanism, further ensuring transparency and integrity. This smart contract-based access control builds off role-based access control, allowing for more complex hierarchies to be used. This model aims to allow for both modularity, making adoption easier for existing digital forensic tools, and encouraging digital investigation and litigation to become more streamlined. Existing tools can easily integrate with the proposed model adding an extra layer of non-repudiation, transparency, and integrity.

Keywords: Blockchain, Chain of custody, Smart contract, Access control, Integrity, Digital forensics

1. Introduction

In the realm of digital data and information, the concepts of digital evidence and integrity stand as paramount pillars, ensuring the reliability and trustworthiness of electronic information in an ever-evolving digital landscape. Digital evidence encompasses data or information stored or transmitted in binary form, serving as a credible and admissible source of evidential information in a court of law (International Organization for Standardization, 2015). Examples of digital evidence include log files, text files, images, and emails. This digital data becomes difficult to preserve owing to no inherent measures to prove its legitimacy (Casey, 2002). A hash function is not sufficient alone as it can only show the file has not been modified since the calculation of the hash (Rivest, 1992).

A strict process is followed when a digital investigation is conducted and is referred to as the Digital Forensic Life Cycle (International Organization for Standardization, 2012). The steps taken include Acquisition, Identification, Collection, Report, and Preservation. During the acquisition and reporting processes, every interaction with a piece of evidence is documented such that the analysis findings can be repeatable and verifiable (Stoykova, 2023). The Chain of Custody (CoC) is a document that tracks all changes in any facet of the investigation. This may include the evidence itself or investigator logs generated by digital forensic software. The integrity of this log must be maintained in order for the evidence to be admissible in a court of law (Prayudi and Sn, 2015). If this log is altered or tampered with in any way, the entire investigation could be put in jeopardy (Sadiku, Shadare and Musa, 2017). The preservation process of digital evidence in regards to storing the chain of custody becomes extremely difficult (Giova, 2011; Prayudi and Sn, 2015). Monitoring the chain of custody during a digital investigation can be quite challenging when attempting to use the standard paper-based approach (Giova, 2011). Therefore, a more reliable and less error-prone approach is to have a digital-based chain of custody. However, this brings challenges regarding its integrity such as ensuring the digital data that was acquired during acquisition did not change during the investigation and was not maliciously modified. This is due to the fact digital data is much easier to modify and can be done remotely with minimal traceability (Cosic and Baca, 2010; Lone and Mir, 2019; Prayudi and Sn, 2015). Hence, a novel model becomes imperative for the safeguarding of chain of custody integrity. Implementing a digital CoC introduces its own unique set of challenges, as there will be a constant need for updates, making it infeasible to securely store a hash in a fixed location, as the hash naturally updates with the ongoing updates to the CoC throughout the investigative process.

Blockchain maintains CoC integrity through digitally signed transactions grouped into blocks, generating a hash. Changes invalidate the chain, causing a snowball effect (Nakamoto, 2008). Existing research (Bonomi, Casini and Ciccotelli, 2018; Burri, et al., 2020; Lone and Mir, 2019; Tian, et al., 2019; Yan, et al., 2020) utilized blockchain as a method for CoC immutability, however, it lacks built-in access control mechanisms as well as integration into other tools. This paper developed a model that allows for more intricate access control mechanisms with the integration of modular storage engines for digital evidence.

2. Related Work

Digital evidence management and access control are core aspects of any digital investigation and several models and strategies have been researched in the field (Chopade, et al., 2019; Lone and Mir, 2019; Singh, Ikuesan and Venter, 2022; 2019). It is evident that blockchain remains the best method for transparency and immutability, however, due to the intricacies of blockchain and the complex technical knowledge and proper use of the technology it becomes poorly adopted. The work by Bonomi, Casini and Ciccotelli (2018), proposed a blockchain-based chain of custody framework, utilizing a private or permissionless blockchain. The solution contains evidence logs, a front end, and a confidential evidence database distributed across trusted entities. One drawback of the solution is the computational overhead when accessing the CoC, as it requires all transactions of the blockchain to be queried. Moreover, a significant issue arises when considering access control, as each individual piece of evidence is solely linked to a single owner, which makes it extremely difficult to work as a team. One possible approach might include distributing the private key among various team members, however distributing a private key is highly risky and can easily result in the key falling into the wrong hands.

Burri, et al. (2020) propose a mixed-chain approach that uses both private and public chains, eliminating the need for a dedicated frontend. A trusted party manages the private chain, while a public chain is used for verification and transparency. This technique creates anchor points, allowing verification of blocks before they are uploaded to the public chain. However, this approach raises integrity concerns as it allows for modification of the private chain without affecting the public chain's immutability and verification.

Yan, et al. (2020) propose a shift from a singular owner model to a revocable ciphertext policy attribute-based encryption, allowing multiple people to collaborate on the same case. This method ensures blockchain integrity and non-repudiation, enhancing teamwork. However, this approach may expose the door for a 51% attack due to the use of attribute-based access control. A hybrid system with a group signature scheme and role-based access control could address this issue. This paper treats each piece of evidence as a transaction, not a block, so each transaction is chained. This type of system makes consensus procedures like mining difficult and computationally expensive because the nonce must be calculated per transaction, not per block.

Current studies have demonstrated the potential of blockchain technology in maintaining the integrity of the CoC. However, there exists a requirement for advanced access control methods that enable collaboration among investigative teams, as well as a faster way to build a CoC. There is also a need for more measures allowing for the user to interact directly with external sub-systems along with the blockchain such as integration with external storage systems.

3. IntegriStore

The proposed model ensures the integrity and non-repudiation of both the chain of custody along digital evidence. It is split into three (3) main components, namely Blockchain, Storage Engine, and a Web Interface as shown in Figure 1. The Blockchain provides the needed integrity, transparency, and non-repudiation, while also containing the information to build and maintain the CoC. The Storage Engine is responsible for the actual physical storage of the digital evidence. The Web interface provides a layer of abstraction for the user and allows them to interact with both the blockchain and the storage engine from a single interface.

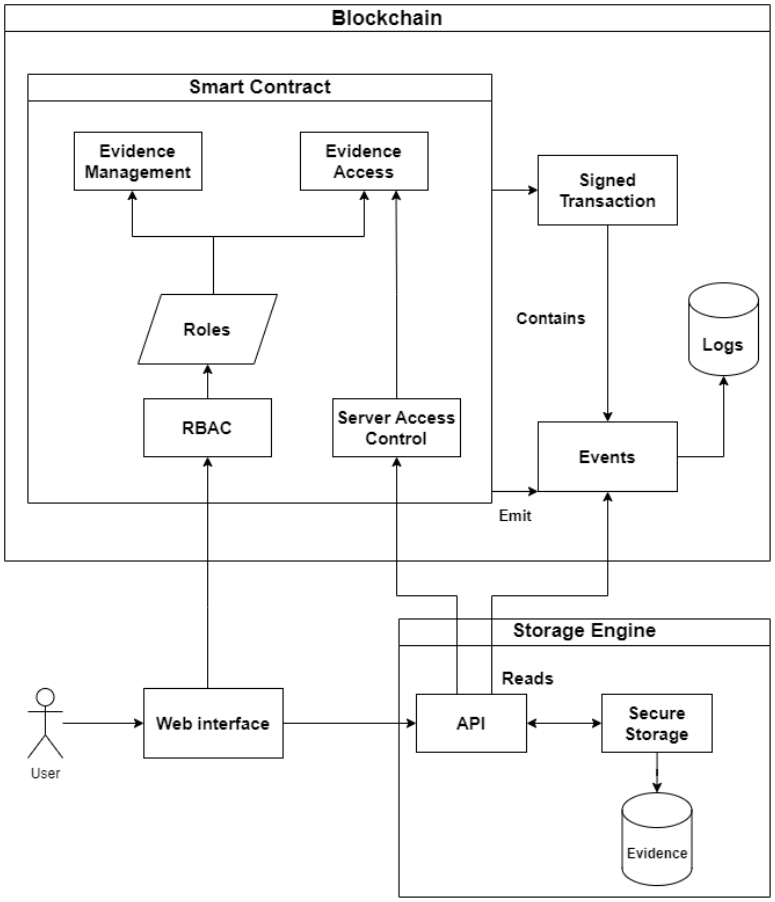


Figure 1: High-level model

3.1 Blockchain

The core of the model is the blockchain, ensuring the integrity of the evidence through cryptographically linked blocks. The consortium blockchain was chosen out of the 4 most common blockchain types (Private, Public, Consortium, and Hybrid (Arooj, Farooq and Umer, 2022)). A public chain uses distributed nodes across members of the public, private is hosted by a single entity, hybrid is a mixture of public and private, and lastly consortium deploys the nodes across trusted parties. Using a public chain is infeasible due to the monetary cost required to execute transactions on the blockchain. Finally, the system would be unsuitable because of the wait times for blocks to be mined on public chains (Doubloin, 2023). A private chain cannot be used since it would significantly reduce the integrity of the system. The hybrid chain suffers from the same cost aspect as the public chain, leaving the consortium blockchain as the best option.

In this case, trusted parties are government institutions such as courthouses and federal buildings. While distributing the chain can increase the risk of someone maliciously getting a copy, it would be difficult to parse as all information is in hash form with no context.

3.1.1 Smart contract

The smart contract handles the role-based access control, the evidence management, server access control and evidence access. Having the contract responsible for such a load places strain on the blockchain. However, this is done so that the chain of custody logging becomes inevitable. In contrast, regular systems can have instances where information and changes do not get logged. Every investigation should have a dedicated smart contract, thereby enabling distinct roles and evidence management for each investigation. This approach introduces a layer of encapsulation, where each investigation is encapsulated within its dedicated smart contract. Even after an investigation concludes and the associated evidence is removed from the storage engine, the smart contract perseveres as a lasting record. This then serves as a permanent history of that investigation which uses minimal storage since evidence is not stored on the chain.

3.1.2 Role based access control

The contract is built by extending the Open Zeppelin Role Based Access Control (RBAC) Scheme (OpenZeppelin, 2023), which inherits and adds more functionality to the access control scheme. In this scheme each role has an admin role, where a role cannot be the admin of itself. Multiple roles can, however, have the same admin, creating a tree-like structure as seen in Figure 2. The root structure of a system terminates with "0x0", and no user can hold this role. A root role is the root of their branch and can add other root roles, but they have no control over them. A user can hold multiple roles, but they cannot hold a role that is the admin of another role they have. A user can assign users to sub-roles but cannot be added to root roles. Any role can be deleted, and once removed, it cannot be added back. This is to ensure proper logging and data integrity. In IntegriStore, integrity and confidentiality can be maintained if a root role is compromised. The storage engine can ignore calls using the compromised contract, and a new contract with new root roles can be made. The integrity is still maintained implicitly, since the logs are immutable, and no one can delete or modify them. The evidence, therefore, is far more likely to be accepted in a court, as the trust in the CoC remains fully intact.

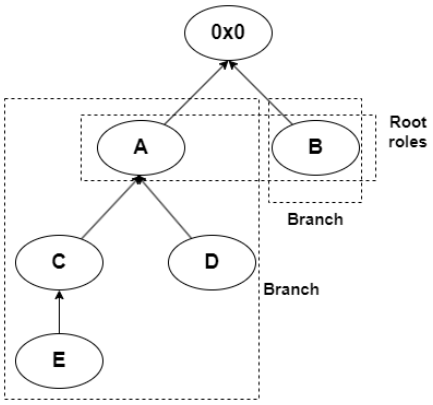


Figure 2: Role

3.1.3 Evidence management

For efficient evidence management, groups were implemented, allowing for the bundling of multiple pieces of evidence together. The smart contract keeps track of which roles have access to which groups. A user that holds a role that has access to a group, can upload an evidence hash to the smart contract. This will trigger an event that acts as the log for the chain of custody. An event tracks who added the evidence, the group it was added to, and the evidence itself in the form of a SHA256 hash. This is the only form of evidence stored on the chain. No other information about the evidence apart from its hash is stored. The contract is not designed to replace the storage/metadata of evidence, rather it is designed to assist the storage engine in adding a layer of immutability. A piece of evidence cannot be “removed” from the contract. If any changes need to be made or the wrong piece of evidence is added, then a new add call needs to be created which will result in a new event. The system is designed to be completely immutable, as any removal or change capability would hinder the credibility of the investigation and logs. A system of giving the user the ability to move evidence as well as remove evidence while logging these changes, may be possible, but is not explored in this paper.

3.1.4 Server access control

In instances where the external storage engine needs to make requests and transactions to the blockchain, the normal RBAC scheme will not work due to the storage engine having different actions from the user. To solve this, the contract keeps a list of addresses that are supported “servers”, i.e. addresses that are trusted servers that may perform server-related actions. This list can be updated, however, only root admins can add/remove these servers from the contract. One example of an operation that only a server can perform, is calling the *saveEvidence* function. This function indicates a particular piece of evidence has been saved to the database of a storage engine. The necessity for the storage engine to call the contract to provide a confirmation upon saving a piece of evidence, stems from the two-stage process of evidence storage. Initially, the hash is uploaded to the blockchain, and only once this is completed will the server accept and securely store the corresponding evidence. This deliberate two-stage design guarantees the comprehensive and accurate logging of all pieces of evidence within the chain of custody.

3.1.5 Evidence access

Since the actual evidence is stored off-chain in a storage engine, the investigator will need to interface with this storage engine to download and upload evidence. If the storage engine were to only use its internal access control mechanisms, it would defeat the entire purpose of the smart contract. In the case of downloading / uploading evidence, the smart contract determines if the user has access to a piece of evidence or the ability to upload it to a particular group. This gives rise to a minor challenge concerning how a user provides verification to the storage engine. To resolve this issue, the "personal_sign" feature of MetaMask is used (Consensys, 2023). To guarantee that the request of a user to store evidence is logged, the user must first make a call to the contract to request for evidence, after which the process as highlighted in Figure 3 is then carried out. The unique hash, generated for each request along with the *Evidence Access* event, prevents against any replay attacks.

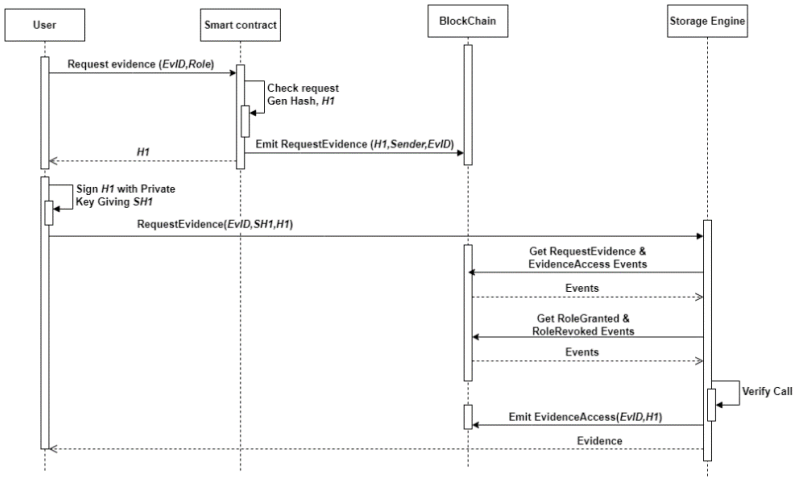


Figure 3: Sequence diagram for evidence access

3.1.6 Events

An event is an immutable log that is attached to a signed transaction when it is emitted from a smart contract. The event log (along with the transactions) implicitly becomes the CoC. Figure 4 shows an example of a decoded event on the blockchain, using Ganache (ConsenSys Software Inc, 2022). This event log is immutable as it is attached to the transaction. It is comprised of a signature (A in Figure 4.), and the parameters (C-D in Figure 4.). If the parameters are indexed, it means that the blockchain can be queried to only return events that match a particular filter. This could be used, for example, to get all pieces of evidence uploaded by a particular person. While log provision is one of the primary functions of these events, they serve another pivotal role. Every alteration made to the contract's state is associated with a prescribed event, allowing the event history to effectively encapsulate and describe the current system state. This allows the contract state to be calculated by replaying the events in chronological order. The API will use these events to verify a request from a user during the evidence access and upload stage as shown in Figure 3.

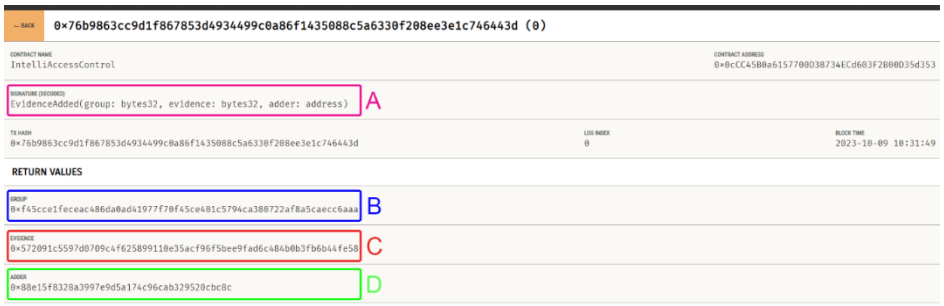


Figure 4: Example event on Ganache

3.1.7 Signed transaction

Each call to the contract is essentially a signed transaction that is sent to the blockchain. If the transaction is verified, it gets accepted and put into a block, making it secure and immutable. Since the user must sign this

transaction with their private key it means that each transaction in the blockchain can be traced back to an address allowing for the needed aspect of non-repudiation.

3.2 Web Interface

To improve adoption and assist the user in interacting with the smart contract, a web interface was utilized. A web interface removes the need of a user to install any application as well as improving device compatibility. The web interface is required to conceal the complexity and intricacies of dealing with the smart contract functions. It is responsible for abstracting the connection between blockchain and user, and user and API. Abstracting into a single interface, allows the user to focus solely on the investigation while the web interface manages the underlying complexity of simultaneously dealing with the 2 sub-systems (Blockchain and Storage Engine). It is crucial that the web interface allow the user to interact with the blockchain directly. The alternative would require the user to pass their private key along, which could negatively impact the safety of the system.

3.3 Storage Engine

The storage engine is responsible for the physical storage of evidence. Storing the evidence on a chain would be too costly. Storing the evidence in any highly distributed manner creates issues of privacy. Highly distributed storage works extremely well for its use case, however, when dealing with investigations where data is extremely sensitive and potentially voluminous, it becomes a breach of privacy. The storage engine, therefore, needs to resemble a highly secure centralized system with redundancy. The access control of this system is achieved through the blockchain, as described in Figure 3. The storage engine does not verify the user from a username and password. All verification is achieved through the private key of the user, the events, and the smart contract. To further improve security, more access control can be added by means of Multi-Factor Authentication. This paper does not define a particular storage engine. It is designed to be modular and integrate with any storage engine, provided the base API functionality to interface with the smart contract is present. Although trusted parties receive access to the blockchain, this does not imply that they also have authority over the storage engine. The blockchain may be hosted by one entity, and the storage engine by another. This makes it possible for a single blockchain to increase in efficiency and integrity, while supporting various storage engines installed or used by digital investigators. However, the storage engine also needs to have its own processes for evidence storage, as well as privacy preservation. For this existing secure storage models such as SecureRS (Singh, Ikuesan and Venter, 2022) and Digital Witness (Nieto, Roman and Lopez, 2016) can be used.

4. Prototype

The prototype was developed using Ganache, an Ethereum-based blockchain, to facilitate the development and deployment of smart contracts. The contract was developed using tools like Hardhat (Nomic Foundation, 2023) and Truffle (ConsenSys Software Inc, 2022), with a solidity-based contract. The web interface, API, and storage engine were built on a single Python flask (Ronacher, 2023) server, with the connection between the user and blockchain carried out client-side through Web3js (ChainSafe, 2023) and EthersJs. MetaMask was used as a wallet provider and for 'personal_sign' features. A straightforward interface was designed using HTML and Bootstrap to hide the complexities associated with both systems from the user. This paper uses the Flask server's local filesystem for file management, allowing direct access control through the API. Web3.py (Ethereum Foundation, 2023) is used for communication between the API and the blockchain. The API is streamlined, with a simplified database containing hashes and their corresponding keys. This approach enhances usability by storing role and group names as hashes on the blockchain, providing a more contextual understanding of data.

Since the storage engine is not the focus of this paper as it is designed to be interchangeable, the local filesystem of the Flask server was used. This allowed the file management to be done directly through the API with only one extra layer of abstraction, which was the access control of the API. The API itself is built into the flask server. Web3.py (Ethereum Foundation, 2023b) was used for the communication between the API and the blockchain. The API did not make any transactions itself, instead it uses Web3.py (Ethereum Foundation, 2023b) to get events from the blockchain. Web3.py (Ethereum Foundation, 2023b) provides the ability to query the blockchain for all events that contain a particular parameter or match a filter. For example, these events were used for the “Verify call” section of Figure 3. This design choice facilitated the creation of a highly streamlined API with a simplified database containing hashes and their corresponding keys (names). This approach greatly enhanced usability, as both the role and group names are stored solely as hashes on the blockchain. Consequently, the system can seamlessly replace the hash with the associated string, to provide users with a more contextual understanding of the data.

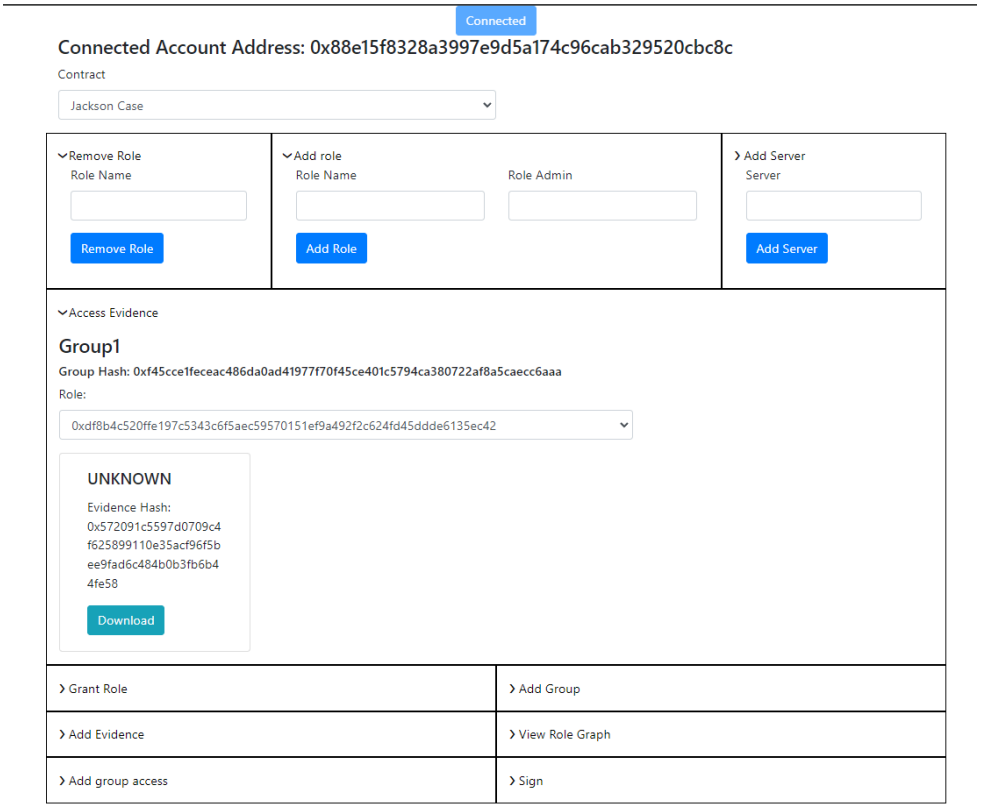


Figure 5: Web Interface

5. Discussion

To benchmark the proposed model and prototype, a comparative analysis was carried out comparing existing research. From Table 1, it is observed that IntegriStore is feature rich and is the only model that allows for role-based access control and contract access control. It is also suggested that the model is unique in allowing for the handling of access control directly through a smart contract, forcing it to be logged directly onto the CoC. IntegriStore is a model that allows for the integrity and non-repudiation of any form of digital data. Since the contract only keeps track of the hash and not the metadata, there is no limitation on the type of data. Any form of data that can produce a hash, may be tracked with the contract. This implies that the model has the capability to store and process more than just digital evidence. One example scenario involves using the model to control physical access control, wherein access cards employing a private key, are utilized to authenticate individuals. In this context, rather than downloading evidence, the focus lies on gaining entry to an evidence room. The model is built around its modularity and its ability to be attached to existing storage engines. Different digital forensic investigations and accompanying tools may differ in the way that they are conducted. The role of the digital forensic investigator and the tools necessary to manage the system, will evolve over time. IntegriStore is designed to evolve and adapt with necessity and change. The storage engine that is required to work with the system is considered external to the actual model, allowing it to be swapped out for whatever tool is needed by the user. It is also possible for the storage engine to be swapped out during an investigation. This would involve taking the current storage engine offline and moving the evidence to a new secure storage engine. After the move, the new storage engine should report the hashes of the evidence it received to the blockchain where it can be ensured no evidence was altered or removed. This allows for evidence integrity and preservation to be maintained when evidence is moved to a new storage engine. This is useful if a vulnerability is detected in the current storage engine as it can be swapped out while the CoC and Evidence Hashes remain secure on the blockchain.

Table 1: Comparative Analysis of IntegriStore

Feature	IntegriStore	(Bonomi et al., 2018)	(Yan et al., 2020)	(Burri et al., 2020)	(Santamaría et al., 2023)
Evidence Integrity	✓	✓	✓	✓	✓
Chain of custody Integrity	✓	✓	✓	✓	✓
Nonrepudiation	✓	✓	✓	✓	✓
Role-based access control	✓	X	X	X	X
Group evidence Ownership	✓	X	✓	✓	X
Contract Access control	✓	X	X	X	X
Multi Key signature	X	X	✓	X	X
Smart contract	✓	✓	X	X	✓
Storage Engine Integration	✓	X	X	✓	✓

6. Conclusion and Future Work

This study introduces an innovative model designed to ensure the integrity and preservation of digital evidence and the chain of custody, ultimately increasing the likelihood of digital evidence being deemed admissible in a court of law. The foundation of this model relies on blockchain and smart contract technology, where IntegriStore comes into play. IntegriStore incorporates smart contract-driven role-based access control, eliminating the constraint of a single owner per piece of evidence. This, in turn, fosters improved collaboration among team members during digital forensic investigations. Furthermore, by entrusting access control to the smart contract, the logging of the chain of custody becomes guaranteed. What is particularly advantageous is that the system can read the blockchain state directly from indexed events, allowing for seamless integration with external subsystems without burdening the blockchain with additional overhead. The web interface offers users direct interaction with both the blockchain and the storage engine, simplifying the overall experience. However, a notable limitation of this model pertains to the potential complexity and size of the smart contract, which may present challenges in terms of management and navigation. For future work, it is recommended that efforts be directed toward creating a more streamlined, smart contract structure. This would help optimize the management of the blockchain while maintaining the effectiveness of the model. A proof of concept has been developed to showcase IntegriStore's feasibility and its ability to preserve the chain of custody, providing a strong foundation for future refinements and expansion of the model. Future research could explore ways to make the model more scalable and adaptable to different use cases, potentially offering a valuable tool for the digital forensic investigators' community.

References

Arooj, A., Farooq, M.S., and Umer, T., 2022. Unfolding the blockchain era: Timeline, evolution, types and real-world applications. Journal of Network and Computer Applications. 207, 103511. https://doi.org/10.1016/j.jnca.2022.103511

Bonomi, S., Casini, M., and Ciccotelli, C., 2018. B-coc: A blockchain-based chain of custody for evidences management in digital forensics. ArXiv Prepr. ArXiv180710359. - [online] Available at: <https://drops.dagstuhl.de/storage/01oasics/oasics-vol071-tokenomics2019/OASIcs.Tokenomics.2019.12/OASIcs.Tokenomics.2019.12.pdf>

Burri, X., Casey, E., Bollé, T., and Jaquet-Chiffelle, D.-O., 2020. Chronological independently verifiable electronic chain of custody ledger using blockchain technology. Forensic Science International: Digital Investigation. 33, 300976. https://doi.org/10.1016/j.fsidi.2020.300976

Casey, E., 2002. Error, uncertainty and loss in digital evidence. International Journal of Digital Evidence . vol 1.

ChainSafe, 2023. web3.js - Ethereum JavaScript API — web3.js 1.0.0 documentation [online] Available at: <https://web3js.readthedocs.io/en/v1.10.0/> [Accessed 20 October 2023].

Chopade, M., Khan, S., Shaikh, U., and Pawar, R., 2019. Digital forensics: Maintaining chain of custody using blockchain. In:IEEE, *Third international conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).* Palladam, India, pp. 744–747. https://doi.org/10.1109/I-SMAC47947.2019.9032693

Consensys, 2023. The crypto wallet for Defi, Web3 Dapps and NFTs | MetaMask [online] Available at: <https://metamask.io/> [Accessed 20 October 2023]10.20.23).

ConsenSys Software Inc, 2022. Ganache - Truffle Suite [online] Available at : <https://trufflesuite.com/ganache/> [Accessed 20 October 2023].

Cosic, J. and Baca, M., 2010. A framework to (im) prove" chain of custody" in digital investigation process. In: *Central European conference on information and intelligent systems.* Faculty of Organization and Informatics Varazdin, p. 435.

Doubloin, 2023. How long will an Ethereum (ETH) transaction stay pending? [online] Available at: <https://www.doubloin.com/learn/how-long-ethereum-eth-transaction-pending> [Accessed 18 October 2023].

Ethereum Foundation, 2023a. Contracts — Solidity 0.4.24 documentation [online] Available at: <https://docs.soliditylang.org/en/v0.4.24/contracts.html#events> [Accessed 19 October 2023]

Ethereum Foundation, 2023b. Web3.py [online] Available at: <https://web3py.readthedocs.io/en/stable/> [Accessed 20 October 2023].

Giova, G., 2011. Improving chain of custody in forensic investigation of electronic digital systems 11. International Journal of Computer Science and Network Security, VOL. 11 No. 1, January 2011

Lone, A.H. and Mir, R.N., 2019. Forensic-chain: Blockchain based digital forensics chain of custody with PoC in Hyperledger Composer. *Digital. Investigation.* 28, 44–55.

Nakamoto, S., 2008. Bitcoin: A peer-to-peer electronic cash system.- [online] - Available at: < https://bitcoin.org/bitcoin.pdf>

Nieto, A., Roman, R. and Lopez, J., 2016. Digital witness: Safeguarding digital evidence by using secure architectures in personal devices. *IEEE Networks.* 30, 34–41. https://doi.org/10.1109/MNET.2016.1600087NM

Nomic Foundation, 2023. Hardhat | Ethereum development environment for professionals. [online] Available at: <https://hardhat.org/> [Accessed 20 October 2023].

OpenZeppelin, 2023. Access Control - OpenZeppelin Docs. - [online] Available at :< https://docs.openzeppelin.com/>

Prayudi, Y. and Sn, A., 2015. Digital chain of custody: State of the art. International Journal of Computers and Applications114, 1–9. https://doi.org/10.5120/19971-1856

Rivest, R., 1992. The MD5 message-digest algorithm. - [online] - Available at: < https://www.rfc-editor.org/rfc/rfc1321>

Ronacher, A., 2023. Welcome to Flask — Flask Documentation (3.0.x) [online] Available at: <https://flask.palletsprojects.com/en/3.0.x/> [Accessed 20 October 2023].

Sadiku, M.N.O., Shadare, A.E. and Musa, S.M., 2017. Digital chain of custody. International Journal of Advanced Research In Computer Science and Software Engineering. 7, 117. https://doi.org/10.23956/ijarcsse.v7i7.109

Santamaría, P., Tobarra, L., Pastor-Vargas, R. and Robles-Gómez, A., 2023. Smart contracts for managing the chain-of-custody of digital evidence: A practical case of study. *Smart Cities* 6, 709–727. https://doi.org/10.3390/smartcities6020034

Singh, A., Ikuesan, A.R. and Venter, H.S., 2019. Digital forensic readiness framework for ransomware investigation. In: F. Breitinger and I. Baggili, eds. *Digital forensics and cyber crime, lecture notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. Cham: Springer International Publishing. pp. 91–105. https://doi.org/10.1007/978-3-030-05487-8_5

Singh, A., Ikuesan, R.A. and Venter, H., 2022. Secure storage model for digital forensic readiness. *IEEE Access* 10, 19469–19480. https://doi.org/10.1109/ACCESS.2022.3151403

Stoykova, R., 2023. The right to a fair trial as a conceptual framework for digital evidence rules in criminal investigations. *Computer Law Secure* Review. 49, 105801. https://doi.org/10.1016/j.clsr.2023.105801

International Organization for Standardization, 2015. *ISO27043, Information technology — Security techniques — Incident investigation principles and processes.*

International Organization for Standardization, 2012. *ISO27037, Information technology — Security techniques — Guidelines for identification, collection, acquisition and preservation of digital evidence.*

Tian, Z., Li, M., Qiu, M., Sun, Y. and Su, S., 2019. Block-DEF: A secure digital evidence framework using blockchain. *Information Science*. 491, 151–165. https://doi.org/10.1016/j.ins.2019.04.011

Yan, W., Shen, J., Cao, Z. and Dong, X., 2020. Blockchain based digital evidence chain of custody. In: *Proceedings of the 2020 2nd international conference on blockchain technology*.Hilo HI, ACM. https://doi.org/10.1145/3390566.3391690

Cyber Resilience, Dependability and Security

Angela Mison[1], Gareth Davies[2] and Peter Eden[1]
[1]University of South Wales, Pontypridd, UK
[2]University of the West of England, Bristol, UK

angela.mison@southwales.ac.uk
Gareth13.Davies@uwe.ac.uk (corresponding author)
peter.eden@southwales.ac.uk

Abstract: There is a continuing skills shortage associated with digital security and DevSecOps (World Economic Forum, 2023), but this paper argues that is due to non-recognition that it is time for cyber security and/or digital security to be defined, and a further separation of specialisms in computing to be made apparent. This has become increasingly important when considering Artificial Intelligence. The problem is not new. This paper presents a refinement of the principles suggested by Milner (2007) of using a model to describe behaviour and organise software, grappling with seemingly intractable and complex problems which cross boundaries between different systems: engineering, technological, social, economic, legal, and political, each with a distinct perspective and goal. It emphasises Hoare's (1996) assertion that system failures are largely due to failed analysis impacting development of resilient systems. It argues that there are dichotomies between resilience – a system security/safety perspective, dependability – a user/consumer perspective, and security – a technology perspective. Many proposed systems to date have conflated these perspectives in the secure by design paradigm which requires a depth of knowledge and expertise. Unicorns are rare. This paper suggests how to overcome the skills shortage utilising the skill sets that are available in a manner that maximises the contribution to digital security. Recognising that not everyone and everything needs to communicate with the world reduces complexity and can increase trust. Concentration on the operational purpose of a system, resulting in an Operational Design Domain (ODD) reduces complexity further. Additional reduction in complexity is achieved by placing resilience in an engineering and programming development context, grounded in acceptable behaviours, while accepting dependability as a user expectation of system behaviour, and cyber security as a separate specialism addressing access to systems and infrastructure. Much of this paper is a reversion to defensive programming through the ODD. There is a need for any solution to the skills shortage be scalable and economic, and this paper suggests how that can be achieved using existing skill sets targeted at their specialisms.

Keywords: Cyber security, Resilience, Dependability, Digital forensics

1. Introduction

There is a myriad of definitions of cyber security, itself, part of the problem. The top Google search engine returns for a definition of cybersecurity demonstrates the subtle differences in emphasis and understanding of the term.

From the NCSC's (2023) reduction of risk, itgovernance.co.uk's (2023) the application of technologies, processes and controls, CISA's (2023) 'protection of networks devices and data from unauthorised access or criminal use and the practice of ensuring confidentiality, integrity and availability of information', to Kaspersky's (2023) 'cyber security is the practice of defending computers, servers, mobile devices, electronic systems, networks, and data from malicious attacks', the definitions highlight the priority of the technology element of organisations concerned and/or their starting point.

The paradigm of security by design (His Majesty's Government, 2023) is an attempt to solve the cyber security issues no matter the priority or starting point indicated above. The Cybersecurity and Infrastructure Security Agency (CISA) is making a series of recommendations, attempting to transfer the majority of the responsibility for software from the end user to the provider (Riotta, 2023). The issues are complicated by the ubiquitous nature of systems, the promotion of hyperconnectivity as a business benefit and the use of Agile technologies to maximise those hypothesised benefits (McKinsey, 2022).

Something very few wish to address is the building of new systems which interact with or depend on legacy systems, where the knowledge and expertise for those systems has long since departed and the documentation may be inadequate. Legacy systems are likely to perform a fundamental function within an organisation, which is both an advantage and disadvantage . The fact that they continue to function can leads to lack of investment until circumstances combine to produce a catastrophic meltdown (Charette, 2014). The identified benefit of retaining legacy systems can be that their very historical nature can renders them arcane and in some cases, unknown to even insiders, through contemporary unfamiliarity with the logic, languages and technologies supporting those legacy systems (King, 2022). Hitherto, malefactors seem to have ignored critical infrastructure legacy systems, looking for easy pickings and higher returns, but now, they are viewed as a less secure and a simpler way to gain access to more secure systems with returns due to ransomware (Raywood, 2023).

To place any system in its operational context, the term digital security is equally relevant and reflects the impact of systems, their use on a daily basis, and the dependence of society on their availability and integrity, either directly or indirectly. Perhaps, having the broadest perspective, the Organisation for Economic Cooperation and Development (OECD) (2023) has an apt definition of digital security and takes an economic and societal interpretation, that extends cyber security to consider impact. The OECD definition includes "the economic and social aspects of cybersecurity, as opposed to purely technical aspects, … 'digital' is consistent with expressions such as digital economy, digital transformation, and digital technologies. It forms a basis for constructive international dialogue between stakeholders seeking to foster trust." This definition is the basis of creating the ODD and its operational boundaries. The ODD is derived from the fundamental understanding of the purpose and expected behaviour of a system.

There are other broader perspectives of cyber security, and this includes those who control budgets and those concerned with the societal impacts attributable to failures of traditional cyber security. As an example, having the ears of Boards and Executives, cyber security can be seen as "not just about managing risk, it's also a strategic issue that shapes product capability, organizational effectiveness, and customer relationships, … identifying where the business creates value and analysing threats, … [working] to de-risk enterprise platforms, extract value from existing investments, secure value chains, and embed 'security by design' into new products and businesses" (Mckinsey.com, 2023). From the McKinsey statement it can be seen that consequences of a successful cyber attack may represent existential risks to organisations .

Further difficulties are caused by the lack of understanding and distinguishing between the ill-defined concepts of information warfare and cyber warfare. The two combine in digital warfare, whether the protagonists are nation states, organised crime groups, guerilla crime groups using Crime as a Service, or individuals demonstrating their capability. Everything in cyberspace is fair game. Access is the objective. What comes after that is opportunistic.

2. An Alternative Approach to Addressing the Skills Shortage

It may seem strange to suggest that the solution to the skills shortage in cyber security could be addressed by revisiting exactly how systems are developed, maintained, and updated. Taking advantage of the existing skills and strengths of the existing personnel, can result in the matching of those skills and strengths appropriate to the lifecycle phase (Paulsen & Byers, 2019). A common vocabulary for qualities of systems is needed (US Food and Drug Administration, 2023). The establishment of common vocabulary lends itself to the matching of which skill sets are required, where and when. Doing this, ensuring common understanding, enables the complex task of achieving maximum cyber security to be achieved, particularly when discussions cross skills boundaries

Part of the problems associated with the skills shortage is the overwhelming nature of imposing 'cyber security' over an organisation and its systems, and the experience and knowledge of detail required by those responsible for their defence.

Simplifying the problems, analysis, to just resilience and security results in the realisation that there are organisational operational / functional applications, based on the purpose of the organisation, and the infrastructure, which becomes the responsibility of cyber security and includes operating systems, standard applications eg word processing, spreadsheets, etc, networks and devices. They can be identified through recognising what system or application is organisation specific and requires knowledge of that data that is moving through the system, the underlying processes, and to whom the data should be communicated. This process definition assists in the establishment of virtual networks and identification of endpoints which become the responsibility of cyber security.

Secure by design (His Majesty's Government, 2023) remains aspirational. It requires everyone involved from the Board to the end user to be able to understand an ill-defined, nebulous, and multiway interpretable security when they are doing whatever task is allocated. This implies a knowledge of both the system and cybersecurity requirements that they may not have. Even more telling, is the impossibility of ensuring they are all aware of the requirements of digital security.

More applications are being developed that contain an element of intelligence. In advance of mass applications being developed, the mantra of 'Responsible AI by design' paradigm (Lu et al., 2023), demonstrates the future difficulties for cybersecurity and digital security as ever more complicated and complex than current circumstances due to machine learning, evolution and the requirement for both audit and transparency of algorithms. This is not only for the organisation but extends to the requirement for forensic explainability should a system ever be associated with an adverse event or decision.

While this paper suggests a possible approach to addressing the skills shortage, at this point, it is worth introducing the caveat in regard to AI. If the AI system is to evolve while in operation, it is not necessary to gain access to the internals of the system to corrupt it, merely to influence its learning. At this time, no solution is completely foolproof in protecting against adversarial machine learning. In the same way that misinformation and disinformation can affect human thinking and learning, the same is true for AI. That systems are open to system on system psy ops is the province of the military industrial complex.

However bad the skills shortage is now, it is likely to get worse. How much more knowledge is expected of these omniscient unicorns? Current practice is linked with extending their expertise rather than on concentrating on what is within their skill set and, for a change, enabling them to contribute to the delivery of a system that works and meets its objectives in a way that is judged as successful. According to Gartner (Panetta, 2021), "48% of CIOs and technology executives having deployed or planning to deploy [AI] technologies in the next 12 months."

Figure 1 suggests the skills split between business functions, systems development, and cyber security. It should be noted that the behaviour required to achieve business objectives is defined through process definition, leaving the developer to exercise the skills of coding the process to achieve the required behaviour. Securing the Application Program Interface (API) of organisation specific systems also occurs within development as the parameter sets within the ODD become known.

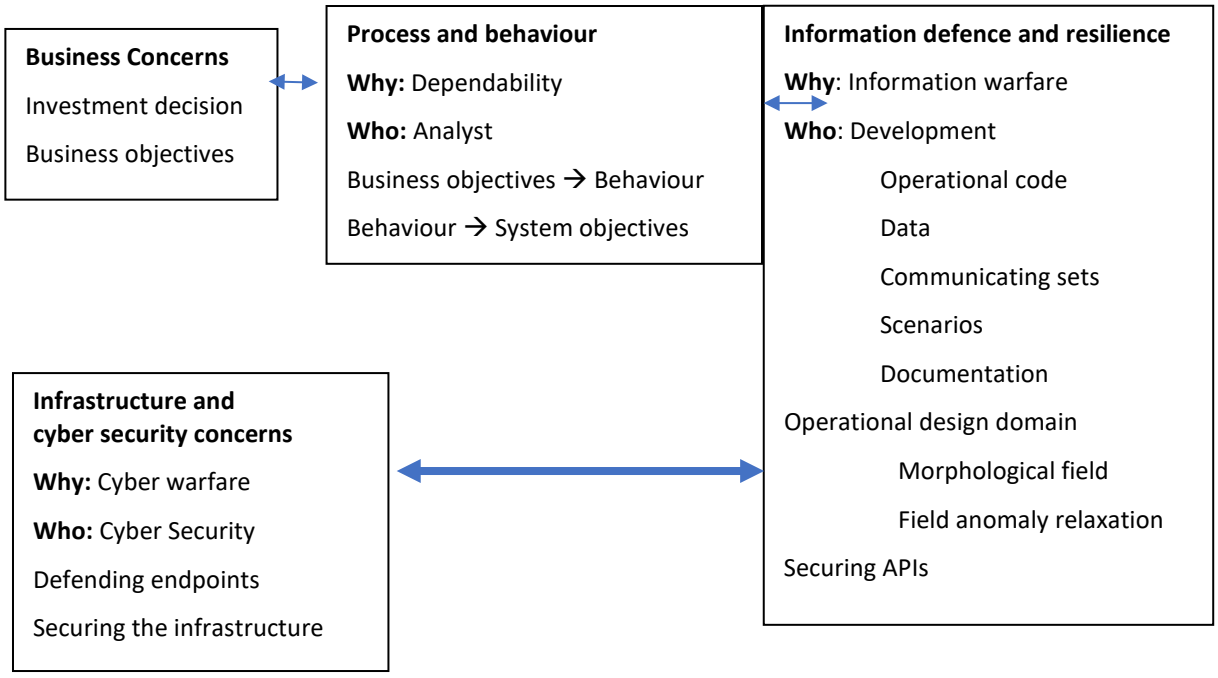


Figure 1: The who, what and why of digital defence, distinguishing necessary skill sets

3. Dependability, Analysis, and the Expected Behaviour of a System

Dependability is defined as "the quality of being able to be trusted and being very likely to do what people expect" (Cambridge English Corpus, 2023). This relates to the behaviour of a system.

Reduction of complexity argues that dependability is an end-user perspective which should be covered through analysis. Satisfaction of this results in a system that works according to everyone's or everything's expectation. Even if incorrect parameters or timings for execution are presented to the system under consideration, the system should continue to behave, in an expected manner. It may do so in a reduced capacity and/or capability, or degrade gracefully but will, otherwise, behave in an expected manner or advise the end-user of an error to avoid resistentialism (Deutscher, 1956).

It is unlikely that many considered the consequences of the UK's National Air Traffic Services (NATS) reaction to an inconsistency and safety concern to be acceptable behaviour. The cause was a single aircraft's flight details which caused the NATS system to shut down and move to a backup system which could remain operational for four hours. NATS quickly realised that it would take longer than the time available and transitioned to manual

input of flight plans. While there is sympathy for NATS being subject to Finagles Law of Dynamic Negatives (techopedia, 2021) on a peak travelling day of the year, there must be approval of their safety considerations. No one wants a midair collision. The repercussions in the broader context were financially staggering and took a minimum of three days to resolve due to positioning and capacity issues (Hand, 2023). To demonstrate where safety is an ostensible imperative, but where limitation of scenario development was even more catastrophic, a Cruise autonomous vehicle operating in San Francisco was involved in a collision with a pedestrian. The process defined that the vehicle should move to the side of the road to prevent a build up of traffic. Sensible, but the unfortunate pedestrian was lying, undetected, under the vehicle (Brodkin, 2023) which was not subject to manual intervention.

Analysis has a dual aspect, defining system objectives and defining behaviour contributing to the achievement of those objectives. The objectives are set by the organisation, whether it is increasing functionality or maximising the benefits of surveillance capitalism. The question for the analyst is not 'What is the system going to do?', but 'Why has the organisation committed to this investment and what does it expect to achieve?'. For an end-user, the behaviour of the system, linked to process definition, as in 'Who has to do what?' will either achieve or contribute to the objectives defined through the behaviour of the end-user, independent of this being a human or sub-system, in response to system demands. Dependability can be seen as predictability for the end-user in completing any system-based task. Where the end-user is a sub-system, the expected inputs and outputs should be defined in that outputs contribute or act as inputs to the next step and should be dependable in their own right.

The analysis is where systems often fail. As identified by Hoare (1996), "Programs have now got very large and very critical ... There have been many problems and failures, but these have nearly always been attributable to inadequate analysis of requirements or inadequate management control." Too often analysts are faced with a situation in which protagonists define a 'normal or standard operation' for a process, but then tack on the phrase, 'but, sometimes….'. The 'sometimes' are, in reality, extensions to the base case and more properly the subject of scenario definition and development as they impact the operational design domain (ODD), a "set of operating conditions under which a given … system … or feature thereof is specifically designed to function" (ISO/TC 204 Intelligent Transport Systems, 2022). It should be recognised that any system development has its origins in an objective whether it is to make the system more secure or efficient, or to enable evolution or expansion of it, or an organisation, some way. This makes it an organisational project which contributes to the objectives of the organisation. That project requires iterative analysis in which the sponsors of the project need to be committed and have their time allocated to the project.

4. Resilience, Safety, and System Development

Resilience is defined as "the quality of being able to return quickly to a previous good condition after problems" (Cambridge English Corpus, 2023). Resilience is covered in development in collaboration with analysts in respect of behaviours resulting from specific scenarios outside the core ODD as defined by the initial analysis.

The initial analysis defines the system in its optimum working state - a standard process with no variation and everything working as defined. The element of failure of analysis arises here, in the variation of system operation to accommodate scenarios which, deriving from field anomaly relaxation, due to lack of iterative analysis input. Habitually there is a single analysis, and a contributing lack of commitment from commissioning entities.

When presented with an incomplete first pass analysis, developers may make logical decisions based on insufficient input which may contribute to the failure of the system when deployed. However intuitive and logical a decision may appear to a developer, in any deployed system, it may be completely counterintuitive to the ultimate end-user or even, the organisation, thereby marring its dependability through resultant unexpected behaviour. Murphy's Law (2012) applies. Further analysis and commitment from the commissioning entity may be necessary during the definition of scenarios which in effect relate to the development of contours within the ODD from mild variation to incipient failure.

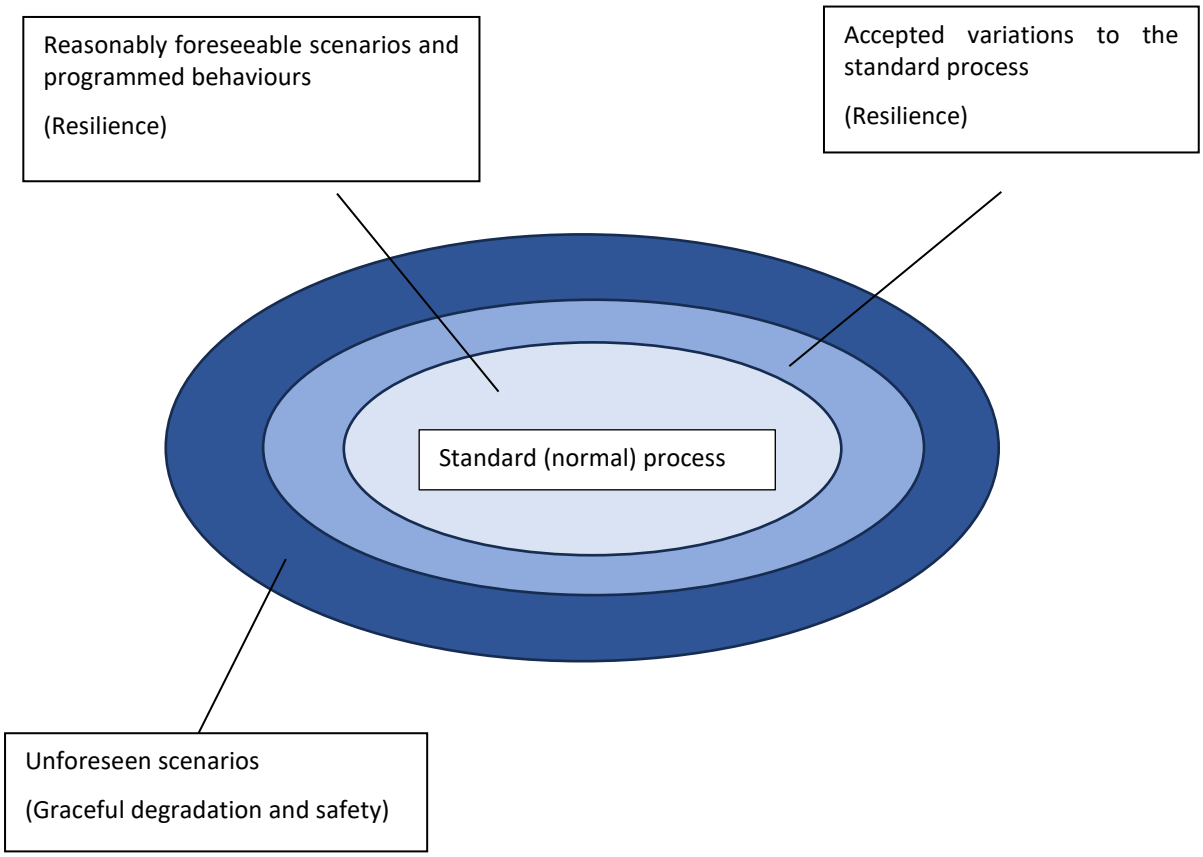


Figure 2: Example of operational design domain contours

Referring to the NATS failure, it is disconcerting to be in a chaotic situation, as has happened historically in some airspaces, where pilots continually need to broadcast their positions in an effort to avoid mid-air collisions and provide the air traffic information to others in the air. The solution to the NATS failure is simple in retrospect. It involves a pre-screening of flight data and the rejection of those not in compliance with expectations of the ODD, either acceptable parameters, or failure due to other errors. They are not presented to the system proper. Rather than the entire system transitioning to manual operation, it is the rejected flight data that is dealt with manually while all other flights continue to be processed automatically. This type of event or scenario has been stated, by NATS, as having an extremely low probability which would place it in the outermost reaches of the ODD contours unless pre-screened. The logic of determining behaviours attributable to the ODD Contours is limited by the developers' and analysts' imaginations of the 'What if …' situations, as demonstrated by the Cruise autonomous vehicle.

In terms of scenario definition, this is to put the system into the real world where acceptable, random and unpredictable acts happen. It also entails using the imaginative and creative powers of future gazing by developers and analysts. It is useful to consider scenarios and decision making as based on game theory and statistics, which have sound mathematical foundations although it should be recognised that use of Bayesian statistics may result in inconsistencies and non-determinism as probabilities are calculated to include historical choices. The end result is a lot of defensive programming which, with luck, will be used rarely.

In terms of defining operational design contours, these are dependent on scenarios relevant to the system under consideration, their probability of occurrence, and priority. There can be as many as necessary, and each contour can be subject to testing, or used to define the test suite, thereby ensuring that at each level the system works behaviourally as defined in the analysis. Defining test suites in line with operational design contours enables the test team to simplify the view of the system and ensure that a complete testing of the system is carried out.

As an example, a new user, despite any training, may attempt to input incorrect data. The engineering practices of Field Anomaly Relaxation (Nazmiye, et al., 2020) to determine scenarios, and/or creating a Morphological

Field, a component of general Morphological Analysis (Ritchey, 2022), can form the basis of judgement and audit as to resultant behaviour, and risk minimisation, lending itself to the future requirement of explainability of the system and its behaviour, in forensic circumstances. This becomes even more relevant and important when AI is added to the mix.

5. Information Defence, Data Flows and Intra-System Communication

The climate of information warfare exists and although the concepts are well understood, through laissez faire or lack of capability, organisations and systems continue to be subject to exploitation of vulnerabilities (Stupples, 2015). Whether it is organised by nation states, organised criminal groups, unethical competitors, or as guerilla tactics, it is for developers to implement API security to combat information warfare and for cyber security to implement defence mechanisms against cyber warfare for the infrastructure.

In addition to the elements of defining the ODD and its contours, information defence becomes part of business continuity. Developers are aware of who should be communicating with whom through the formation of virtual networks from their associated communicating sets. The ODD provides the list of acceptable operational parameters, or parameters for which there is a programmed resolution and behaviour.

Under these conditions, it becomes possible for developers to consider the failure of cyber defence in which a malefactor has gained access to the organisation's operational systems. The developers can include algorithms at the API, which identify rogue parameters putting the systems under threat. These rogue parameters may include source, destination, timing, or input or output parameters outside those within the ODD contours. These algorithms can initiate isolation of source and automatically advise cyber security of a potential breach and discovery location – removal of the domino in the race which halts a race in its tracks. It should be noted that the term 'isolation' is used, in this paper, to refer to a system or component part of a system rendering it unable to communicate or interact with other parts of the system.

Information security is built around the priorities determined by the organisation, addressing issues proactively rather than reactively as currently occurs. This can be done only in the knowledge of the system, its objectives, and its context of operation. It derives from the concept of 'defending' cyber assets (Aguiar, 2023), (NIST Computer Security Resource Center, 2023), as opposed to cyber infrastructure.

6. Cyber Defence, Endpoint Security and Segmentation

Future attacks and attackers cannot be known. A system must be defended continuously, against all malefactors and errors, and notify its defenders of the system location when it has been breached or subject to an attempted breach. For this reason, the collaboration between the developers who have documented the data flows and necessary communicating parties in any process provides the knowledge of the system, its priorities and weak points to those responsible for the overall defence of the organisation's digital assets. Depending on the level of security required, data flows, the responsibility of developers, can be validated at the API through resilience initiatives and use of isolation, while access to the network can be controlled by cyber security.

Where there is a need for universal access, as in random customer access to a retail website, reference can be made to the behavioural characteristics expected of the end-user, the customer. Where this deviates from the prescribed behaviour, as happens in many fintech applications, a limited number of attempts can be acceptable, or the end user can be locked out and a procedure usually requiring human customer/customer service interaction/intervention, must be completed before the lock is removed.

It can be seen from the foregoing, that digital access can be treated in much the same way as resilience through the definition of scenarios and access contours. The worst case scenario being segmentation of the external network through exclusion of all communications both to and from an offending party. Here, the slight repurposing of a Morphological Field would indicate the addresses and functions affected, which in turn relates to the potential isolation of a component of a system as it becomes non-functioning in the absence of available data.

In cases where on-going communication between parties is essential, back up lines of communication, person-to-person, can be triggered. Such communication, if technology based would require an alternative mechanism for communication, such as a land line, or satellite phone. The matching of the cyber defence contours together with the information defence contours defines, completely, the requirement and priority for the initialisation of business continuity procedures.

These arrangements are fully within the expertise of cyber defence, and do not require detailed knowledge of organisation specific systems beyond that. It demonstrates the importance of interactions with other areas of expertise and knowledge, but these interactions are between similarly educated and linguistically inclined personnel, although emerging from their education into different spheres.

7. Conclusion

There is increasing dependence on ubiquity and availability of systems, from society as a whole, down to individual organisation. Oftentimes, the concept of system unavailability seems completely alien and beyond comprehension. Historically, systems development and budgets have been pared to the bone. Simply put and referring back to Hoare (1996), secure systems cost more in the initial stages, which can be protracted.

A major contributing factor to losing track of systems and increasing organisational vulnerability, both in terms of interconnection and infrastructure, is the time lag relating to documentation resulting from the Agile, digital transformation push promoted by reputable management consultancies. Boards believe that Agile development is both cheap and quick, an essential component to the much promoted monetisation and efficiencies available from digital transformation and hyperconnectivity. Someone in IT and cyber security needs to convey to the Board that systems will cost more, in sheer documentation terms, if they are to be as secure as possible.

As a caveat, there is an expectation of the continuous availability of cyber defence staff and incident response teams. Reality can intervene. The current political situation in the United States of America may well provide an opportunity for its traditional and non-traditional enemies to launch cyber attacks while politicians argue over budgets and there is a potential for government shutdown. The implications of a shutdown, for any military in action, are more than significant, and potentially disastrous. This is a policy issue outside the control of any cyber defence unit. For national security, it is elements such as this that define the operational context and provide asymmetric advantage to adversaries.

References

Aguiar, A., 2023. *Defining Assets in Cybersecurity Asset Management.* [Online] Available at: https://noeticcyber.com/defining-cyber-asset-management/[Accessed 8 December 2923].

Brodkin, J., 2023. *After robotaxi dragged pedestrian 20 feet, Cruise founder and CEO resigns.* [Online] Available at: https://arstechnica.com/tech-policy/2023/11/after-robotaxi-dragged-pedestrian-20-feet-cruise-founder-and-ceo-resigns/[Accessed 8 December 2023].

Cambridge English Corpus, 2023. *Cambridge Dictionary.* [Online] Available at: https://dictionary.cambridge.org/dictionary/english/dependability[Accessed 25 September 2023].

Charette, R., 2014. *RBS Group Facing Huge Fine over Massive 2012 IT System Meltdown . Bank still fixing decades-long neglect if IT system infrastructure.* [Online] Available at: https://spectrum.ieee.org/royal-bank-of-scotland-group-facing-huge-fine-over-2012-massive-it-system-meltdown[Accessed 6 December 2023].

Cybersecurity & Infrastructure Security Agency (CISA), 2023. *What is cybersecurity.* [Online] Available at: https://www.cisa.gov/news-events/news/what-cybersecurity

Deutscher, I., 1956. Physicians' Reactions to a Mailed Questionnaire: A Study in 'Resistentialism.. *The Public Opinion Quarterly,* 20(3), pp. 599-604.

Hand, J., 2023. *Nats air traffic control: Experts reflect on three days of chaos.* [Online] Available at: https://www.bbc.co.uk/news/uk-66685349[Accessed 25 September 2023].

His Majesty's Government, 2023. *Secure by design.* [Online] Available at: https://www.gov.uk/government/collections/secure-by-design[Accessed 25 September 2023].

Hoare, C., 1996. *Unification of Theories: A challenge for Computing Science.* s.l., Springer Verlag, pp. 49-57.

itgovernance, 2023. *What is Cyber Security? Definition and Best Practices.* [Online] Available at: https://www.itgovernance.co.uk/what-is-cybersecurity

Kaspersky, 2023. *What is cyber security?.* [Online] Available at: https://www.kaspersky.com/resource-center/definitions/what-is-cyber-security[Accessed 25 September 2023].

King, A., 2022. *https://www.securityindustry.org/2022/10/14/legacy-systems-rip-and-replace-or-keep-them-going/.* [Online] Available at: https://www.securityindustry.org/2022/10/14/legacy-systems-rip-and-replace-or-keep-them-going/[Accessed 8 December 2023].

Lu, Q., Zhu, L., Xu, X. & Whittle, J., 2023. Responsible-AI-by-Design: A Pattern collection for designing responsible artificial intelligence systems. *IEEE Software,* 40(3), pp. 63-71.

Mckinsey.com, 2023. *Cybersecurity.* [Online] Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/mckinsey-technology/overview/cybersecurity [Accessed 25 September 2023].

McKinsey, 2022. *The data-driven enterprise of 2025.* [Online] Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025 [Accessed 15 March 2022].

Milner, R., 2011. *Transcription of the Presentation Is Informatics a Science.* [Online] Available at: https://markstaples.com/files/Is%20Informatics%20a%20Science%20-%20v1.0.pdf[Accessed 27 April 2023].

Murphy, C. E., 2012. *Murphy's Laws Origin.* [Online] Available at: https://militaryhumor.net/murphys-laws-origin/[Accessed 9 September 2021].

National Cyber Security Centre (NCSC), 2023. *What is cyber security.* [Online] Available at: https://www.ncsc.gov.uk/section/about-ncsc/what-is-cyber-security

National Cyber Security Centre, 2019. *Secure design principles.* [Online] Available at: https://www.ncsc.gov.uk/collection/cyber-security-design-principles[Accessed 24 October 2022].

NIST Computer Security Resource Center, 2023. *asset definitions.* [Online] Available at: https://csrc.nist.gov/glossary/term/asset[Accessed 8 December 2023].

OECD, 2022. *Digital Security.* [Online] Available at: https://www.oecd.org/digital/digital-security/[Accessed 8 December 2023].

Organisation for Economic Cooperation and Development (OECD), 2023. *Digital Security.* [Online] Available at: https://www.oecd.org/digital/digital-security/[Accessed 25 September 2023].

Panetta, K., 2021. *IT Budgets are Growing. Here's Where the Money's Going.* [Online] Available at: https://www.gartner.com/en/articles/it-budgets-are-growing-here-s-where-the-money-s-going[Accessed 25 September 2023].

Paulsen, C. & Byers, R., 2019. *NIST IR 7298 Rev. 3 Glossary of Key Information Security Terms.* [Online] Available at: https://doi.org/10.6028/NIST.IR.7298r3 [Accessed 8 December 2023].

Raywood, D., 2023. *What are the biggest threats to OT?.* [Online] Available at: https://www.infosecurityeurope.com/en-gb/blog/threat-vectors/operational-technologies-biggest-threats.html[Accessed 8 December 2023].

Riotta, C., 2023. *The nation's cyber defense agency is continuing to drive a major effort to shift security responsibilities from users to software providers.* [Online] Available at: https://www.nextgov.com/cybersecurity/2023/09/cisa-plans-new-secure-design-guidance/390019/[Accessed 25 September 2023].

Ritchey, T., 2022. General Morphological Analysis: An overview. *Academia Letters*, January, p. Article Number 4620.

Stupples, D., 2015. *What is information warfare?.* [Online] Available at: https://www.weforum.org/agenda/2015/12/what-is-information-warfare[Accessed 25 September 2023].

techopedia, 2021. *Finagle's Law.* [Online] Available at: https://images.techopedia.com/definition/19342/finagles-law[Accessed 9 September 2021].

US Food and Drug Administration, 2023. *Glossary of Computer System Software Development Terminology (8/95).* [Online] Available at: https://www.fda.gov/inspections-compliance-enforcement-and-criminal-investigations/inspection-guides/glossary-computer-system-software-development-terminology-895[Accessed 8 December 2023].

World Economic Forum, 2023. *The Future of Jobs 2023,* Geneva: World Economic Forum.

Cross-Disciplinary AI Supply Chain Risk Assessment

Angela Mison[1], Gareth Davies[2] and Richard Ward[1]
[1]University of South Wales, Pontypridd, UK
[2]University of the West of England, Bristol, UK

angela.mison@southwales.ac.uk
Gareth13.Davies@uwe.ac.uk (corresponding author)
richard.ward@southwales.ac.uk

Abstract: While AI remains chip based and part of both commercial and national strategic superiority goals, it is useful to examine the security and risks associated with achieving those goals. The future strategy rests perilously on an unstable inverted triangle of financial and economic reality. This paper presents the AI chip supply chain as an inverted triangle which base/apex is dependent on a global single supplier with the capability of producing equipment essential for their manufacture. It highlights the dependence on a single company for the fabrication of those chips, and the security risks associated with that supplier being Taiwanese in limited foreign ownership. It is suggested that the increasing tensions between China and the USA have resulted, in part, from this dependence, which was demonstrated by the supply chain crisis resulting from Covid-19. The attempt to reduce this dependence led to the CHIPS and Science Act 2022, signed into law by President Biden. In part of the inverted triangle are found Big Tech and the major Cloud Service Providers. They vary between 60% - 80% of their market capital being in financial institutional ownership, most of which is held by a very limited number of institutions, not all of whom are publicly quoted. To doubt the influence wielded by those financial institutions, just a single institution with major Big Tech and Cloud holdings has, at 31 December 2022, USD 8.59 trillion of assets under management. This represents economic power and places it between the equivalent Gross Domestic Product of China (USD 19.37 trillion) and Japan (USD 4.41 trillion) the second and third entries behind the USA in the GDP rankings. Financial institutions are market driven to achieve growth, contribute to economic stability, and are to an extent regulated by unelected vested interests and organisations. The battlefield for national supremacy of AI may concern chips, until the arrival of quantum AI. Current Chinese economic woes are providing the momentum for pre-emptive strikes at the semiconductor industry, and an inverted triangle is neither a secure nor stable structure for a supply chain.

Keywords: AI, Aladdin, GAMMA, TSMC, ASML

1. Background

Chips for Artificial Intelligence (AI) are the most powerful chips manufactured. AI systems' learning and decision making require them to process vast quantities of data obtained through historical datasets and surveillance capitalism platforms. Currently, the fabrication process for these chips requires extreme ultraviolet (EUV) lithography to mass produce patterns on silicon. EUV lithography systems provide the highest resolution in high volume manufacturing (ASML, 2022).

The principles associated with EUV lithography can be thought of as similar to layering photographic prints of up to 100 layers on a single chip. Of the EUV lithography machines currently in existence, these operate at 13.5nm (ASML, 2023a), but research and development are ongoing. It is intended that High-NA EUV lithography which will advance Moore's Law towards the logic 2nm technology generation will go on pilot release in 2023 before full scale production in 2025 (De Simone, et al., 2021). Without EUV lithography, AI would be dependent on less powerful chips which would increase its cost in terms of creation and maintenance.

ASML, a Dutch company, has a monopoly in EUV lithography and its underlying technologies. It is the world's unique supplier. In terms of financials, the complexity and cost of EUV lithography machines is such that in the second quarter of 2023, ASML sold 107 new units and 6 used units for a total of USD 6 billion, of which USD 1.7 billion is attributable to EUV lithography units, providing a gross profit margin of 51.3% (ASML, 2023b). ASML's primary customer for EUV lithography machines is Taiwan Semiconductor Manufacturing Company (TSMC).

The United States has negotiated restrictions on sales of ASML EUV lithography systems, in particular to China and its Semiconductor Manufacturing International Corp (SMIC) (Office of Foreign Assets Control, 2021). The US Dept of Commerce designated SMIC as a 'military end user' in October 2020 and applies a 'presumption of denial' for products required to produce chips at 10nm or below.

Some reverse engineering by SMIC, which targets and hires former TSMC employees would appear to have taken place. A lawsuit dated 2004 (Clendenin, 2004) alleged theft of intellectual property from TSMC by SMIC, and was settled in 2005 (Lemon, 2005). In December 2020 the former Head of TSMC R&D, Chiang Shang-yi was appointed Vice-Chairman of SMIC. SMIC is now under investigation for sanctions' violations relating to

provision of advanced 7nm chips for use in Huawei's Mate 60 Pro smartphone, which chips, it is asserted by the US Dept of Commerce, cannot be produced without US technology (Che, 2023). Huawei was placed on the Entity List in May 2019.

TSMC supplies 30% of the world semiconductor market excluding memory output (Taiwan Semiconductor Manufacturing Company, 2023). It is reputed to supply 90% of the AI chip requirement, with Samsung fulfilling the balance. Affecting NVIDIA, AMD, Apple and many others (Shilov, 2023) for the next 18 months, TSMC announced it can satisfy only 80% of demand. This is due to lack of advanced chip packaging services in the manufacturing process which combines a graphics processor unit with six high bandwidth memory chips. Chip packaging provides high speed data transmission and performance that enables training in AI (asia.nikkei.com, 2023). AMD, Intel, and Samsung are potential competitors in this area.

2. AI Development

The discussion on and actual regulation of AI may have slowed down the requirement for TSMC's output. The European Union is the first to regulate AI through its AI Act (News European Parliament, 2023), requiring national supervisory authorities and establishment of a European Artificial Intelligence Board. This publication followed the call for a moratorium on AI citing Risks to Society by over 1,000 technology leaders in March 2023 (Loizos, 2023), although it had probably been in preparation for some time. Nonetheless, with a predicted compound annual growth rate (CAGR) of around 40% (Precedence Research, 2023), far exceeding that of the predicted CAGR of 18% for Cloud (Markets and Markets, 2022), after a period of evaluation of any regulations, it is expected that Big Tech will continue apace with AI development. Indeed, there is a feeding frenzy in stock market as Apple reaches a market Capitalisation value of USD 2.8 trillion (Stock Analysis, 2023).

The predominant end users of the high specification chips produced by TSMC for use in AI systems are anticipated being those Big Tech players, Alphabet (Google), Apple, Amazon, Meta (Facebook), and Microsoft (GAMMA). There will be others, but with for example, Apple having a 23% interest in TSMC production, these are considered to be the majors. Investment strategists suggest that there is a superseding 'Magnificent Seven' which comprises the GAMMA grouping indicated above with the addition of NVIDIA and Tesla (Mitra, 2023). These seven organisations are propping up the Nasdaq with year to date (16 June 2023) stock price increases from between 40% for Apple, Microsoft, Amazon, and Alphabet, to 108% for Tesla, 134% for Meta, and 192% for Nvidia.

In terms of AI development, and for the purposes of this paper, only the GAMMA grouping will be considered further. It is interesting to consider the ownership of the GAMMA grouping by financial institutions (shown in table 1). This is shown in the table below:

Table 1: The percentage of the GAMMA grouping held by financial institutions

GAMMA Group	%
Alphabet (Google)	78.93
Meta (Facebook)	76.35
Microsoft	73.21
Apple	61.32
Amazon	60.53

Further analysis of those financial institutions demonstrates the value of holdings of three of the world's largest financial institutions, Vanguard, BlackRock, and State Street Corp. The holdings shown in table 2 are for those investments directly held by the named institutions and does not include any investments made indirectly by other organisations in which they have holdings. It is likely that the total value of ownership of the three aforementioned financial institutions is obscured by such indirect investment. Like other financial institutions and mutual funds, they have invested and managed a proportion of their assets under management (AUM) in the technology sector.

Table 2: The value of holdings of the GAMMA grouping held by Vanguard, BlackRock, and State Street Corp

GAMMA Group	Market Capitalisation	Ownership by Vanguard		Ownership by BlackRock		Ownership by State Street Corp	
	USD Trillion	USD Billion	%	USD Billion	%	USD Billion	%
Apple	2.799	233	8.38	183	6.62	102	3.69
Microsoft)	2.401	210	8.74	174	7.24	94	3.93
Alphabet (Google)	1.650	63	8.21	54	7.12	28	3.66
Amazon	1.421	100	7.05	84	5.92	46	3.23
Meta (Facebook)	0.784	54	8.14	46	6.88	13	1.98
Totals	**9.055**	**660**	**-**	**541**	**-**	**283**	**-**

With holdings in the GAMMA grouping of this magnitude, financial institutions and mutual funds can directly influence the outcome of shareholder voting and enforce or veto policies or lines of investigation. In general, the financial institutions will have a relationship with the Executive of these companies.

AUM represents the economic power that the financial institutions have. There is a degree of similarity between the economic power accorded to a country through its Gross Domestic Product (GDP) ranking and the value of AUM of a financial institution. The three financial institutions under consideration AUM are shown in table 3.

Table 3: The AUM of Vanguard, BlackRock, and State Street Corp

Financial Institution	Value of Assets
BlackRock	USD 8.5 trillion
Vanguard	USD 7.1 trillion
State Street Corp	USD 3.5 trillion

The AUM of BlackRock and Vanguard places them as an equivalent to the third and fourth national GDP behind USA, between the equivalent GDP China (USD 19.37 trillion) and Japan (USD 4.41 trillion). State Street Corp is not far behind. The influence and economic power of financial institutions and mutual funds cannot be doubted or ignored.

Caution should be exercised when considering these three financial institutions as independent entities. Quite apart from a circular ownership between themselves, each institution uses a model-based risk management platform, Aladdin. Aladdin was originally built within BlackRock to effect monitoring and analysis of the risk profile of their investments.

3. Aladdin

Over the years access to Aladdin has been sold. Aladdin has become a fully-fledged AI system, evolving during the thirty or so years it has been in operation. It now executes approximately 70% of high frequency trades in the US without any human intervention. Aladdin controls approximately 10% of global investments and its expansion into investment of organisations which bulk purchase and outbid individuals in the real estate market is seen as partially responsible for escalation of house prices.

The potential of a 40% CAGR for AI and a total market valuation of USD 196.63 billion in 2023 increasing to a revenue forecast of USD 1.8 trillion in 2030 provides an expectation of growth and income in the intervening years (Grand View Research, 2023a). This is reinforced where the underlying organisation is also a Cloud Service Provider with a 20% CAGR and a total market valuation for Cloud computing of USD 678 billion in 2023 increasing to USD 2.4 trillion in 2030 (Fortune Business Insights, 2023). The figures provided by Grand View Research (2023b) are slightly more conservative.

The use of cloud services by organisations suggests that cloud adoption could generate an additional USD 3 trillion in EBITDA value by 2030 (McKinsey & Company, 2022). This adds indirectly to the attractiveness of investment in technology companies to Aladdin. The calculations for growth and increased revenues both to underlying investments and to the financial institutions themselves makes the logic of investment in these areas undeniable and irrefutable to Aladdin.

The mission statement of BlackRock is assumed to infuse the objectives and behaviour of Aladdin. That mission is to create economic opportunity for all by making investing more accessible. Viewing the AUM figures, it could be said to have achieved its mission, though it is possible that Aladdin is reinforcing particular behaviours of and in the markets.

The head of BlackRock is a member of the trustees of the World Economic Forum (WEF) which espouses a concept of equality in a better world. This concept allies with the original mission statement of BlackRock and is expected to have contributed to the evolution of Aladdin. However, in 2016, a discussion paper was presented to the WEF in which a potential life was described (Auken, 2016). An urbanised population owned nothing, had no privacy and life had never been better. This population lived in a completely service-based economy. The paper also described adverse conditions associated with the existence of people outside the hypothesised utopia. If Aladdin scours the internet in the manner of ChatGPT, it is possible that its original objectives and behaviours have altered, directing its investing decisions to achieve the hypothesised goals of the discussion paper.

'Aladdin is BlackRock's end-to-end investment management and operations platform' ... (BlackRock, 2021a). The Aladdin Data Cloud is a part of Snowflake's Data Cloud and a part of the Aladdin platform (BlackRock, 2021b). Aladdin combines sophisticated risk analytics with comprehensive portfolio management, trading, operations and accounting tools on a single, unified platform. The original goals of Aladdin may have evolved from the primacy of the individual to the primacy of the discussion paper goals. It could be a cautionary tale, validating the suggestion of a moratorium and the need for regulation. Unfortunately, the power and control vested in Aladdin cannot easily be wrested from it without economic and possibly great societal upheaval.

One can ponder the philosophical question of whether the moment has been reached when Aladdin is making investments in AI opportunities because it needs more, better, AI, or according to its original objectives it has analysed AI as a growth area, or its goals and behaviours have determined a better society for humanity and it is working to achieve that.

Having provided the context of the players in the AI supply chain, represented in the inverted triangle below, from ASML to the AI system Aladdin, is it possible to answer the question posed by the title, 'Is the AI supply chain secure, and where and what are the risks?'. As ever, the response is 'it depends', but the indication is that based on the AI supply chain, the future may be unstable, unpredictable and not currently within the control of humanity.

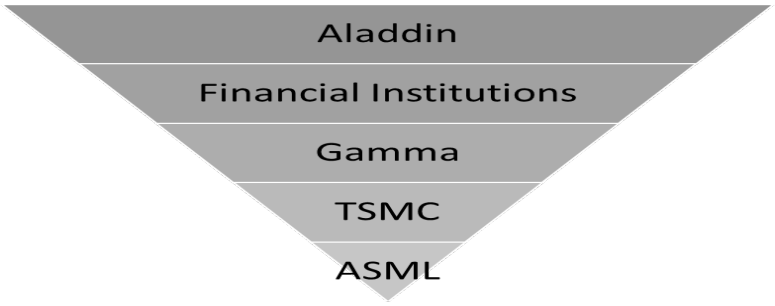


Figure 1: The identified determinants of AI futures

As the world's unique supplier of essential equipment for the manufacture of EUV lithography necessary for use in AI, ASML represents a single point of failure of an economic and scalable AI future until the arrival of quantum computing. Risk is spread due to the number of manufacturing locations for ASML, and the potential for a secondary source of EUV lithography equipment developed and sourced from China.

Currently, AI is largely dependent on the output from TSMC. It has dispersed fabrication plants throughout the world. However, it remains a Taiwanese company subject to the geopolitical forces surrounding the one China policy, the acceptance of which is key to Sino-US relations. The one China policy regards "Taiwan [as] an inalienable part of China's territory … There is no room for compromise or concession" (Wong, 2022).

An ambiguity arises for the USA due to legislation – the Taiwan Relations Act (2022), providing for assisting Taiwan to defend itself, including against a Chinese takeover, and its dependence on TSMC for advanced chips. The USA has singular diplomatic relations with China and not with Taiwan. It is assumed that the CHIPS Act (2022) which endeavours to encourage repatriation of semiconductor manufacture and fabrication within the

USA has resulted both from Covid-disrupted supply chains and the escalation of tension between China and the USA following recognition of its absolute dependence on TSMC.

It is assumed that GAMMA are pursuing the advancement of AI to obtain additional market differentiation. They are both oligopolists and oligopsonists in the technology sector. It has been argued previously that they address differing market segments and are not in direct competition with each other (Mison, et al., 2022), but whether this will remain the case in respect of AI is not known. They are market driven to maintain or grow their positions within the sector. The extent of AI creation and/or adoption, cannot be determined based on existing use cases due to immaturity of the technology and the understanding of it, its capability, and its full potential.

The degree of ownership of GAMMA by financial institutions has been hinted at earlier. A single financial institution, BlackRock, has an equivalent economic power of the world's third ranked GDP. Traditionally, financial institutions are market driven to achieve growth and income via their investment strategies whether that is in research and development or investment. BlackRock is unusual that it has been, for some time, under the control of an AI system. Given the value of AUM, there is a potential for broad economic destabilisation, even global economic instability, should any of the three identified financial institutions fail. To some extent, the institutions themselves and their actions are regulated. What is not regulated is Aladdin, which is fundamental to all three.

The AI system, Aladdin, has been in operation for an extensive period during which time its goals and behaviours may have become modified. It is of concern that as a proprietary system, little is known of the mechanisms of its operation and it is not subject to regulatory control, or audit of its algorithms, objectives, and behaviours. Given the macroeconomic impact of failure, arguably this should be a matter of national security. It cannot be determined if Aladdin is pursuing its own objectives which coincide with those of BlackRock and other parties, or it is on the verge of emergence, the consequences of which cannot be predicted or timed.

The intertwined ownership of financial institutions by other financial institutions is redolent of Enron. It is not surprising the Aladdin arose out of failed ventures but has the philosophy of its human owners that models on which those failures were based could be amended, improved, corrected, which ever word seems appropriate for 'we can build it better'. It is uncertain what influence its creators exercise over the system, but an inverted triangular supply chain with unknown controlling influences seems to provide neither a secure nor a stable basis for the future of AI.

4. Conclusion

So, in response to the question, 'How secure is the AI supply chain?', the answer is likely to remain uncertain until something goes wrong, and its cause can be identified. In the meantime, regulation and caution may be advisable as it has become a national security issue with failure of Aladdin, or the financial institutions having the potential to disrupt the global economy and cause catastrophic societal consequences.

References

asia.nikkei.com (2023) "TSMC sees AI chip output constraints lasting 1.5 years", [Online], https://asia.nikkei.com/Business/Tech/Semiconductirs/TSMC-sees-AI-chip-output-constraints-lasting-1.5-years.

ASML, (2022) "Our products", [Online], https://www.asml.com/en/products.

ASML (2023a) "ASML Annual Report 2022", [Online], https://www.asml.com/-/media/asml/files/investors/financial-results/a-results/2022/2022-annual-report-based-on-us-gaap-jder53.pdf?rev=073f28c1f10641319464651efdad8b3f.

ASML (2023b) "ASML reports EUR6.9 billion net sales and Eur 1.9 billion net income in Q2 2023", [Online], https://www.asmcom/en/news/press-release/2023/q2-2023-financial-result.

Auken, I. (2016) "Welcome to 2030. I own nothing, have no privacy, and life has never been better", [Online], https://medium.com/world-economic-forum/welcome-to-2030-i-own-nothing-have-no-privacy-and-life-has-never-been-better-ee2eed62f710.

BlackRock, 2021a. *BlackRock to launch the 'Aladdin Data Cloud' powered by Snowflake.* [Online] Available at: https://www.blackrock.com/corporate/newsroom/press-releases/article/corporate-one/press-releases/aladdin-data-cloud-powered-by-snowflake

BlackRock, 2021b. *Aladdin Data Cloud.* [Online] Available at: https://www.blackrock.com/aladdin/products/aladdin-data-cloud

Che, P. (2023) "Tech war: Huawei sets September 25 for new product launch but remains tight-lipped in Mate 60 Pro smartphone chip", [Online], https://www.scmp.com/tech/big-tech/article/3234519/tech-war-huawei-sets-

september-25-new-product-launch-remains-tight-lipped-mate-60-pro-smartphone?campaign=3234519&module=perpetual_scroll_2_AI&pgtype=article.
CHIPS Act (2022), [Online], United States Government, https://www.govinfo.gov/content/pkg/PLAW-117publ167/html/PLAW-117publ167.htm.
Clendenin, M. (2004) "*Analysis: background on TSMC, SMIC lawsuit",* [Online], https://www.eetimes.com/alanlysis-background-on-tsmc-smic-lawsuit/.
De Simone, D., Lorusso, G., Philipsen, V. and Ronse, K. (2021) "High-NA EUVL: The next major step in lithography", [Online], https://www.imec-int.com/en/high-na-euvl-the-next-major-step-in-lithography.
Fortune Business Insights (2023) "The global cloud computing market size was valued at $569.31 billion in 2022 & is projected to grow from $677.95 billion in 2023 to $2,432.87 billion by 2030", [Online], https://www.fortunebusinessinsights.com/cloud-computing-market-102697.
Grand View Research (2023a) "Artificial Intelligence Market Size, Share and Trends Analysis Report by Solution, by Technology (Deep Learning, Machine Learning), By End-use, By Region, and Segment Forecasts, 2023 – 2030", [Online], https://www.grandviewresearch.com/industry-analysis/artifical-intelligence-ai-market.
Grand View Research (2023b) "Cloud Computing Market Size, Share & Trends Analysis Report by Service (SaaS, IaaS), by Deployment, by Enterprise Size, by End-User, By Region, and Segment Forecasts, 2023-2030", [Online], https://www.grandviewresearch.com/industry-analysis/cloud-computing-industry.
Lemon, S. (2005) "China's SMIC settles lawsuit with TSMC", [Online], https://www,infowaorld.com/article/2669319/china-s-smic-settles-lawsuits-with-tsmc.html.
Loizos, C. (2023) "1,100+ notable signatories just signed an open letter asking 'all AI labs to immediately pause for at least 6 months'", [Online], https://techcrunch.com/2023/03/28/1100-notable-signatories-just-signed-an-open-letter-asking-all-ai-labs-to-immediately-pause-for-at-least-6-months/.
Markets and Markets (2022) "*Cloud Computing Market by Service Model (IaaS, Paas, and Saas) by Deployment Model, Organisation Size, Vertical (BFSI, retail and consumer goods, telecommunications, IT, ITeS, and manufacturing) and Region - Global Forecast tp 2027",* [Online], https://www.marketsandmarkets.com/Market-Reports/cloud-computing-market-234.html.
McKinsey & Company (2022) "Projecting the global value of cloud: $3 trillion is up for grabs for companies that go beyond adoption", [Online], Available at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/projecting-the-global-value-of-cloud-3-trillion-is-up-for-grabs-for-companies-that-go-beyond-adoption.
Mison, A., Davies, G. and Eden, P. (2022) "Role of Big Tech in Future Cyber Defence", [Online], https://www.researchgate.net/publication/359020725_Role_of_Big_Tech_in_Future_Cyber_Defence.
Mitra, M. (2023) "Forget FAANG, Meet the 'Magnificent Seven' Stocks surging in 2023", [Online], https://www.nasdaq.com/articles/forget-faang-meet-the-magnificent-seven-stocks-surging-in-2023.
News European Parliament (2023) "EU AI Act: first regulation of artificial intelligence", [Online], https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.
Office of Foreign Assets Control (2021) "Sanctions Search List", [Online], https://sanctionssearch.ofac.tres.gov/Details.aspx?id=30951.
Precedence Research (2023) "Artificaial Intelligence (AI) Market Size, Growth, Report by 2032", [Online], https://www.precedenceresearch.com/artificial-intelligence-market.
Shilov, A. (2023) "TSMC controls production of AI processors", [Online], https://www.tomshardware.com/news.tsmc-owns-the-ai-processing-market.
Stock Analysis (2023) "Apple Inc (AAPL)", [Online], https://stockanalysis.com/stocks/aapl/market-cap/.
TSMC (2023) "TSMC Annual Report 2022", [Online], Taiwan Semiconductor Manufacturing Company,: https://investor.tsmc.com/sites/ir/annual-report/2022/2022%20Annual%20Report-E.pdf.
Taiwan Relations Act (2022), [Online], United States Government, Available at: https://www.congress.gov/117/bills/s4428/BILLS-117s4428is.pdf.
The Taiwan Policy Act (2022), [Online], United States Government, https://www.foreign.senate.gov/imo/media/doc/SBS%20Taiwan%20Policy%20Act%20FINAL%20(1).pdf.
Wong, T. (2022) "Biden vows to defend Taiwan in apparent US policy shift", [Online], https://www.bbc.co.uk/news/world-asia-china-61548531.

Proof of Concept of a Digital Forensic Readiness Cybercrime Language as a Service

Maryam Mohamad Al Mahdi and Stacey Baror
Department of Computer Science, University of Pretoria, South Africa

maryam.mohamadalmahdi@up.ac.za
stacey.baror@cs.up.ac.za

Abstract: Welcome to the 21st century, where more devices are connected to the Internet than people on this earth. Each device will be or has been a target of cybercrime. The popularity of text-based cybercrimes has become prevalent with the advancement of the Internet. The number of cybercrimes increases yearly, unlike the reports of such crimes. The major problem is the lack of awareness and protection from cybercrimes. These issues cascade into greater consequences, such as a shortfall of reports, deficiency of cybercrime data, etc. The DFClaaS system is proposed as a solution; it allows users to report text-based cybercrimes anonymously. Next, several NLP techniques are applied to such reports to comprehend text-based cybercrimes as a language, and once the usage of the language is detected, an investigation proceeds. Subscribed users are individuals or businesses that are in some regard connected to the Internet. The DFClaaS system aims to protect its users from text-based cybercrimes and provide digital forensic investigators resources to conduct a successful investigation, enriching digital forensics research. The system is a feature-rich digital forensics readiness tool that will track the aggressive advancements of text-based cybercrimes, thus serving and protecting its users.

Keywords: Digital forensics, Text-based cybercrime, Natural language processing, Phishing, cyberattacks, Digital forensics readiness

1. Introduction

In today's digital age, the Internet has ushered in unprecedented convenience and connectivity, but it has also brought forth a challenge: the proliferation of cybercrime. These illicit activities transcend geographical boundaries and linguistic barriers, posing a pervasive threat in our increasingly interconnected world. The extent and diversity of these cybercrimes are striking, encompassing a wide range of malicious activities, often exploiting the languages we use for communication. Digital forensics, the systematic process of collecting, analysing, and preserving electronic evidence to investigate and prevent cybercrime, plays a crucial role in combatting cybercrime. As we delve deeper into the challenges posed by the proliferation of cybercrime, it is essential to recognize how cybercrimes have evolved, especially in the context of the COVID-19 pandemic.

The COVID-19 pandemic and the widespread adoption of remote work have further exacerbated the cybercrime problem (Chigada and Madzinga, 2021). As society rapidly adapted to digital modes of operation, cybercriminals capitalized on the ensuing vulnerabilities.

The surge in cybercrimes has exposed a troubling trend: many individuals and businesses remain trapped in a reactive posture, scrambling to respond after attacks happen rather than proactively building defenses. This reactive stance breeds a host of critical issues that undermine our collective defense against cyber threats.

First, victims often remain silent, their voices muffled by a complex web of anxieties. Unawareness of reporting channels, fear of embarrassment or blame, concerns about anonymity, and even a lack of faith in law enforcement's ability to act – these factors collectively form a wall of silence that shields cybercriminals from consequences.

Second, businesses prioritize functionality over security, leaving gaping vulnerabilities within their services. This creates a digital landscape fraught with weak authentication protocols, unpatched software, and inadequate data encryption – a veritable buffet for skilled cybercriminals.

Third, the fight against cybercrime is hampered by a lack of robust data. Investigations flounder due to insufficient case information, and the development of crucial machine learning models for predictive analysis suffers from a dearth of training data. This data scarcity weakens our ability to anticipate and thwart future attacks.

Finally, the presence of incompetent third-party services only exacerbates the problem. Lacking the necessary knowledge, resources, and collaborative spirit, these entities delay the identification of cyberattacks and cast doubt on the accuracy of investigative findings. This muddies the waters and allows cybercriminals to operate with impunity.

The current reactive approach towards cybercrime simply is not sustainable. We must shift towards a proactive stance, actively strengthening defenses, fostering open communication, and bolstering our investigative capabilities.

Building upon the insights presented in the study titled "Functional Architectural Design of a Digital Forensic Readiness Cybercrime Language as a Service" (Baror et al., 2023), our research seeks to implement the proposed solution as a response to the surge in cybercrimes. Inspired by the functional architectural design presented in the mentioned paper, our study aims to revolutionize the approach to cybercrime by developing the Digital Forensic Cybercrime Language as a Service (DFClaaS) system. This innovative system, which forms the crux of our research, is designed to address the challenges posed by cybercrimes, particularly those of a text-based nature.

Our methodology involves constructing a high-level model based on the architectural design proposed by Baror et al. (2023), and subsequently implementing a proof-of-concept. Leveraging Natural Language Processing (NLP) models, including Large Language Models, sentiment analysis, and lexicon analysis, to create an efficient and effective tool for combating cybercrimes.

The key objectives include implementing microservices that address various challenges, proposing a system to enhance incident reporting and analysis, and providing valuable resources for investigators. Ultimately, the DFClaaS system aims to benefit many stakeholders, from individual users and organizations seeking a secure digital environment to digital forensic investigators striving to combat the ever-evolving landscape of cybercrime.

The next section follows the background literature on concepts key to understanding this research.

2. Background Literature

In our digitally connected world, text-based cybercrimes have become a pervasive threat, encompassing various illegal activities. As these cybercrimes evolve, digital forensics readiness (DFR) plays a crucial role in responding effectively. DFR involves preparing for digital incidents by developing policies and capabilities to collect and analyse digital evidence. To combat these evolving threats, the integration of cutting-edge technology, such as Natural Language Processing (NLP), is essential. NLP leverages human language understanding to help computers process, analyse, and generate text and speech data, making it valuable in combating text-based cybercrimes. Large Language Models (LLMs) are a significant development in NLP, transforming various fields with their deep understanding of language. This review explores the connections between text-based cybercrimes, DFR, NLP, and LLMs, emphasizing their importance in modern cybersecurity challenges. A review of text-based cybercrime is discussed next to address these emerging challenges.

2.1 Text-Based Cybercrimes

We live in a time of convenience; access to the Internet and devices to connect to it unveils the digital world where almost everyone and everything is accessible with a click of a button. However, this digital age also brings a new frontier of crime, text-based cybercrimes (Meifilina et al., 2019). These crimes encompass a wide range of illegal activities carried out through text communication on the Internet. Examples of such crimes are romance scams, identity theft, cyberbullying, and phishing (Stabek et al., 2010). When text-based cybercrimes are orchestrated, the intention is a win-lose situation. The attacker must gain something such as money or access to personal information, which is effectively what the victim loses (Aftab et al., 2022). Often, these scams appear genuine, making it challenging for individuals to distinguish between legitimate and fraudulent communication.

Text-based cybercrimes represent a significant concern that is often downplayed in terms of its gravity. The pressing need for awareness arises from the firm and the increasing emergence of these deceptive practices, highlighting the importance of recognizing and addressing the prominence of such crimes.

In the ever-evolving domain of cybercrime, where perpetrators exploit the ease of online communication to carry out malicious activities, digital forensics readiness plays a pivotal role in mitigating the impact of these offenses.

2.2 Digital Forensics Readiness

Digital forensics readiness (DFR) refers to the proactive measures taken to prepare for and effectively respond to digital incidents, cybercrimes, and data breaches. It involves the development and implementation of policies, procedures, and technical capabilities to ensure that digital evidence can be collected, preserved, and

analysed in a forensically sound manner (Harbawi and Varol, 2016). Digital forensics readiness is essential in today's technology-driven world to minimize the impact of cyberattacks, enhance incident response, and support legal investigations. Cybercrimes are forever advancing, and the need for digital forensics readiness is a growing concern worldwide. The importance of Digital Forensics Readiness (DFR) cannot be overstated. Its significance lies not just in responding to incidents but in proactively preparing individuals and organizations to safeguard against the pervasive and ever-advancing nature of cybercrimes (Holt et al., 2022).

In the domain of digital forensics readiness (DFR), it's essential to stay one step ahead in the constantly evolving landscape of cyberattacks (Harbawi and Varol, 2016). To achieve this, embracing cutting-edge technology becomes imperative, and Natural Language Processing (NLP) emerges as a pivotal ally.

2.3 Natural Language Processing

Natural Language Processing (NLP) is a field of computer science and artificial intelligence that draws inspiration from human language to enable computers to understand, interpret, and generate it. NLP involves the development of algorithms and models that allow computers to process, analyse, and generate text and speech data in a meaningful and useful way (Chowdhary and Chowdhary, 2020). It leverages lexical and morphological analysis, syntax analysis, semantic analysis, discourse integration, and pragmatic analysis (Chowdhary and Chowdhary, 2020). This multi-faceted approach allows NLP systems to break down the text into individual words, analyse grammatical structure, extract meaning, consider context, and understand implied intentions, facilitating applications like machine translation, sentiment analysis, chatbots, and speech recognition, thereby bridging the gap between human communication and computer understanding. As NLP algorithms continue to evolve, they pave the way for innovative applications. The continuous progress in NLP points towards a future where human-computer interaction is characterised by a deeper understanding and more nuanced interpretation of language (Berger and Packard, 2022).

Within the dynamic field of NLP, a transformative breakthrough has been the emergence of Large Language Models. Large Language Models present a transformative dimension to how we address digital forensics readiness in the context of text-based cybercrimes by offering advanced text analysis insights from textual evidence and improving the identification of cybercrime patterns, ultimately strengthening the readiness and responsiveness to text-based cybercrimes.

2.4 Large Language Models

Large Language Models (LLMs) represent a significant milestone in the field of NLP and artificial intelligence, revolutionizing NLP tasks and applications. These models are meticulously pre-trained on massive datasets containing diverse text from the Internet (Petroni et al., 2019), allowing them to grasp the intricacies of syntax, semantics, and contextual relationships in language. The breakthroughs leading to the emergence of modern LLMs can be primarily attributed to the deep learning revolution with the introduction of neural network architectures. In 2018, models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformer 2 (GPT-2) marked a turning point, showcasing the potential of large-scale, pre-trained neural language models. These models have profoundly impacted a wide range of fields, from chatbots and language translation to text generation and content understanding (Pal et al., 2023), demonstrating their versatility and potential to reshape how humans and machines interact with language.

The following section will provide an exploration of relevant literature in the field of cybercrimes and digital forensics readiness (DFR).

3. Related Literature

Table 1: Related Literature

Referenced literature	Summary of Contents	Evaluation	Conclusion
Ch et al. (2020)	Proposed a generalized framework for classifying cybercrimes, including text-based cybercrimes. The dataset was collected from Kaggle and CERT-In. Achieved 99% accuracy, outperforming some models. The ”others” category lacks specificity. Limited applicability to text-based cybercrimes.	The framework demonstrated high accuracy and outperformed some models, but its limited specificity and focus on Indian cybercrimes may affect its generalization to other countries.	The framework’s generalization for text-based cybercrimes is limited, and future work should focus on more specific categories.
Dinakar et al. (2011)	Introduced a system to recognize textual cyberbullying by scraping YouTube comments. Different labels were used, and models achieved up to 72.96% accuracy. Recognizing sarcasm remained a challenge.	The models showed promising accuracy, especially with specific features, but the challenge of recognizing sarcasm was acknowledged.	The importance of understanding text-based cybercrimes as a language and handling nuances like sarcasm is highlighted.
Janet et al. (2020)	Proposed a system to detect phishing using a balanced dataset of URLs. Achieved 96% accuracy using an ensemble model of CNN and LSTM.	The model’s accuracy was excellent compared to heuristic and list-based methods. LSTM’s data persistence was intelligently utilized for understanding context.	The combination of CNN and LSTM for phishing detection proved highly effective and could be applied in real-world scenarios.
Mahbub and Pardede (2018)	Introduced a system to detect online recruitment fraud through advertisements using the EMSCAD dataset. JRip rule-based classifier achieved the highest accuracy of 96%.	The dataset was relatively small, which may have limited the model’s diversity.	Versatility in the dataset is crucial for developing reliable models for text-based cybercrimes, considering the frequent occurrence and various types of such crimes.

4. Case Study: DFClaaS in Action - Cybercrime Investigation

Characters:

- Alice: A diligent user of the Digital Forensic Readiness Cybercrime as a Service (DFClaaS) system, employed at Bank A.
- Bob: An anonymous cybercriminal attempting to extort Bank A.

Incident: Cyberattack at Bank A
Background:

Bank A, a leading financial institution, faces a severe cyberattack attempt orchestrated by an anonymous threat actor, Bob. Bob demands a substantial sum of money, threatening to release sensitive customer data if the bank fails to comply. Bank A, equipped with DFClaaS, swiftly responds to counter the threat and launch a comprehensive investigation.

DFClaaS Services in Action:

1. Detection and Alert Service:

- As soon as Alice visits the suspicious website associated with the cyberattack attempt, the DFClaaS system activates.

- The system conducts real-time scanning, detecting red flags such as false promises, similarities to known scam websites, etc.
- An alert is triggered, and a pop-up message warns Alice about the potential cybercrime attempt, advising her not to provide personal or financial information.

2. Anonymous Cybercrime Reporting Service:

- Alice, a cybersecurity analyst at Bank A, utilises the anonymous reporting service to report the extortion incident without compromising her identity.
- The reported incident contributes valuable string data to build the DFClaaS database.

3. Reporting Cybercrime Service:

- The system collects reported data about the attempted attack.
- Natural Language Processing (NLP) and machine learning components parse, clean, and analyse the data to identify potential cybercrime patterns as well as generating an analysis report.

4. Cybercrime Text Scraping Service:

- To augment the cybercrime dataset, the system employs web data extraction, scraping relevant string data from external sources.

5. User Management Service:

- Bank A's cybersecurity team, including Alice, interacts with the DFClaaS system.
- The user management service facilitates secure actions, such as logging in, updating profiles, and accessing the DFClaaS system.

6. Visualisation Service:

- Authorised users, including cybersecurity analysts, visualise raw data to identify patterns and potential threats.
- The system's visualisation service enhances usability, enabling a comprehensive analysis of historical string data related to cybercrime language patterns.

7. Cybercrime Text-Data Training Service:

- Deep learning techniques, process the collected cybercrime text data.
- Feature extraction ensures an accurate representation, narrowing down potential cybercrime language triggers.

8. Cybercrime Semantic Builder Service:

- Utilising generated semantic string data, the system builds a cybercrime semantic database.
- Lexical-semantic analysis validates the cybercrime string, creating unique semantics crucial for forensic investigators.

Outcome:

Proactive Cyber Defense:

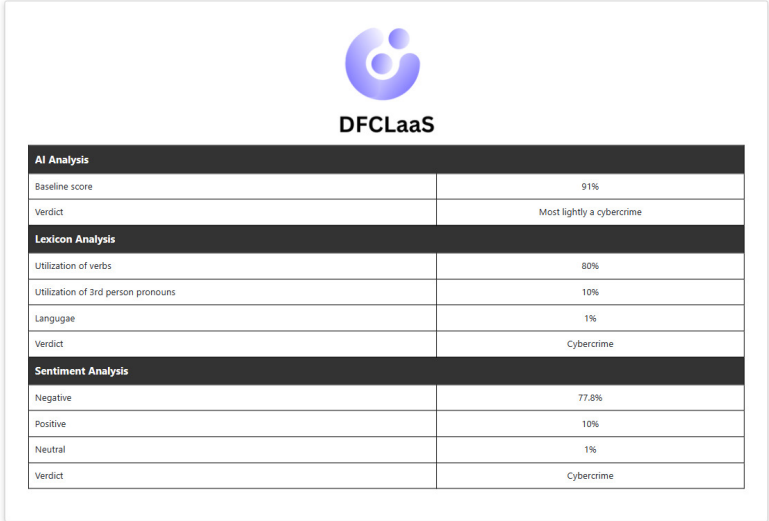
- Bank A's utilisation of DFClaaS ensures a proactive defence against cybercrime attempts.
- The collective services of the system contribute to the early detection, analysis, and prevention of cyber threats.

Secure Digital Forensics:

- DFClaaS provides Bank A with advanced tools for secure digital forensics.
- The system's integration of NLP and deep learning techniques enhances the investigation process, allowing for accurate threat assessment.

Collaborative Cybersecurity:

- The case exemplifies the collaborative nature of DFClaaS, where various services work in harmony to protect against cyber threats.
- Collaboration with a digital forensics investigator ensures a comprehensive analysis and potential legal action against the cybercriminal, Bob.



DFCLaaS

AI Analysis	
Baseline score	91%
Verdict	Most lightly a cybercrime
Lexicon Analysis	
Utilization of verbs	80%
Utilization of 3rd person pronouns	10%
Langugae	1%
Verdict	Cybercrime
Sentiment Analysis	
Negative	77.8%
Positive	10%
Neutral	1%
Verdict	Cybercrime

Figure 1: DFClaaS Analysis Report

In this incident, the DFClaaS system's Detection and Alert Service plays a crucial role in fortifyingBank A's cyber defences by providing real-time warnings and advice to users like Alice, ensuring a proactive response to potential cyber threats.

The next section will offer a proof of concept elaborating on the system's fundamental components and functionality.

5. Proof of Concept

The Proof-of-Concept section delves into the core elements of DFClaaS, shedding light on the crucial components that underpin its functionality. At its heart, DFClaaS is designed to understand and identify text-based cybercrimes. It employs advanced AI and Natural Language Processing (NLP) techniques to assess and analyze the language, lexicon, and sentiment associated with these crimes. In the following subtopics, we will explore the Cybercrime Data Training Service, Lexicon Analysis, and Sentiment Analysis, all of which work in synergy to strengthen DFClaaS's ability to detect and interpret cyber threats in textual content. These integral components are pivotal in digital forensics and cybersecurity, offering a multi-faceted approach to safeguarding users and aiding investigators in their mission.

5.1 Cybercrime Data Training Service

The core of DFClaaS lies in understanding text-based cybercrime in terms of identifying the crime and analyzing the lexicon and sentiment that make up such crimes, essentially the services that implement the various NLP techniques.

The Cybercrime Data Training Service plays a pivotal role in digital forensics and cybercrime detection. This service leverages advanced AI and NLP (Natural Language Processing) techniques to estimate the likelihood of a reported text being a cybercrime. This section delves into the core functionalities and construction of the Cybercrime Data Training Service.

The heart of this service is a meticulously constructed AI model that utilizes the Google BERT model and TensorFlow. The model is designed to process user-submitted reports and assess the probability of the text containing elements of a potential cybercrime. The model construction involves the following essential layers:

- **Input Layer:** The model begins with an input layer that accepts text data.
- **Preprocessing Layer:** The preprocessing layer is responsible for preparing and preprocessing the text data.
- **BERT Encoder:** We utilized the BERT (Bidirectional Encoder Representations from Transformers) model from TensorFlow Hub, a pre-trained language model designed for various natural language understanding tasks.
- **Output Layer:** The model outputs a probability indicating the likelihood of a potential cybercrime.

The Cybercrime Data Training Service serves as the initial line of defense in cybercrime detection. At the same time, Lexicon Analysis further bolsters DFClaaS by enabling the system to comprehend and identify the intricate language associated with cybercrimes within textual content.

5.2 Lexicon Analysis

Lexicon analysis is a critical pillar of the DFClaaS system, serving as a robust tool for identifying and comprehending language related to cybercrimes within textual content. At its core, this analysis hinges on a meticulously constructed lexicon, encompassing a broad spectrum of terms and their corresponding meanings associated with cybercrimes. This lexicon acts as a valuable resource, equipping DFClaaS with the knowledge necessary to detect and interpret the intricate language of cybercrimes.

When a user submits a report, the system initiates lexicon analysis. The attack text within the report transforms into numerical representations through word embeddings, utilizing the GloVe model. This numerical conversion empowers the system to handle the text data in a structured and quantifiable manner. Subsequently, the system employs similarity analysis, specifically using cosine similarity scores, to assess the likeness between the terms in the input text and those present in the lexicon. Higher similarity scores indicate a greater likelihood of cybercrime language in the report.

By recognizing and understanding the nuances of cybercrime language, the system enhances its ability to fulfill its role in the digital forensics and cybersecurity domains. This in-depth analysis equips DFClaaS with the tools required to protect users from cyber threats and offer valuable insights to investigators.

Building on this foundation, Lexicon Analysis equips DFClaaS with the knowledge to detect cybercrime language. Sentiment Analysis enhances the system's detection capabilities by assessing the emotional tone and context within reported content.

5.3 Sentiment Analysis

The DFClaaS system incorporates sentiment analysis to enhance the detection of text-based cybercrimes. Sentiment analysis assesses the emotional tone of reported content, focusing on identifying negativity, neutrality, and positivity. When a user submits a report, this feature comes into play, particularly in evaluating the possibility of a hate crime. DFClaaS combines a profanity checker with a sentiment analyser from the Natural Language Toolkit (NLTK) to achieve this.

The analysis process commences by screening the text for profanity or offensive language, recognizing these as potential indicators of malicious intent. However, the presence of profanity alone is not sufficient to conclude a hate crime. Context plays a crucial role in determining intent. Therefore, the sentiment analyser measures the emotional context within the content, examining negativity, neutrality, and positivity.

The DFClaaS system generates an aggregated score by combining these analyses, providing an indicator of the potential of a hate crime. For instance, high negativity and the presence of profanity lead to an increased likelihood of a hate crime, while a moderate score indicates a moderate likelihood. This approach offers a nuanced assessment that considers both the presence of concerning language and the context in which it is used. Sentiment analysis, thus, contributes to a more accurate and comprehensive detection of potential threats in digital crime.

The results of the different experimental models employed in identifying text-based cybercrimes are discussed in the following section.

6. Results

Before implementing the cybercrime data training service, we conducted experiments with four different models, LSTM autoencoder, Google Palm, Google BERT, and gzip text classification, to determine the most suitable model for identifying text-based cybercrime.

6.1 LSTM Autoencoder

The LSTM autoencoder was initially explored for the task. It was trained on a dataset that included cybercrime and non-cybercrime texts, focusing on anomaly detection. The model's ability to reconstruct input data was used to calculate averaged similarity scores on test data, leading to the determination of a threshold value for identifying anomalies in new attack texts. Unfortunately, the calculated threshold was a value of 269854234.8986, which turned out to be extremely high, considering the closer the value is to zero, the better the accuracy to be observed. This, therefore, indicates a low similarity between input and output texts.

Achieving better results would require a larger dataset and substantial computational resources (GPU and RAM), which were not readily available at the time of this study.

Google Palm was used to try to improve the performance of LSTM. However, this model struggled to process inputs containing inappropriate language, a significant characteristic of text-based cybercrime content. This limitation made the Google Palm model less suitable to achieve the set goals.

6.2 Google BERT

We explored the Google BERT model, known for its pre-training on a vast corpus of internet text data, allowing it to understand language structure and semantics. We fine-tuned the BERT model for the task of classifying attack text, generating a decimal value representing the likelihood of a cybercrime. Validation of the BERT model revealed promising results, with an accuracy rate of 93% and a minimal loss of 28%.

6.3 Gzip Text Classification

A further investigation of gzip text classification involves categorizing compressed text data into predefined classes based on content. The classifier was trained on a dedicated dataset and demonstrated an accuracy rate of 91% and a low loss of 9% when tested.

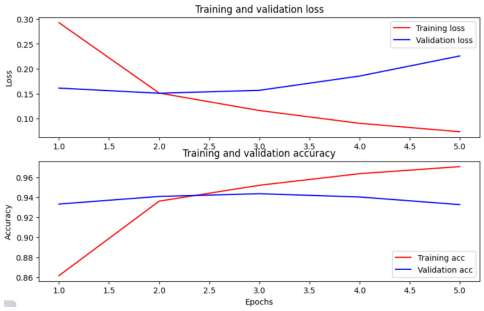


Figure 2: Plot of loss and accuracy during training and validation

These experiments informed the integration of the cybercrime data training service into the DFClaaS system, with the Google BERT model emerging as a key component in our efforts.

The following section will delve into the evaluation of the DFClaaS system.

7. Evaluation

This section delves into evaluating the different models considered for the DFClaaS system. While exploring and evaluating different models, the DFClaaS system ultimately adopted the Google BERT model. While the selection may seem influenced by the model's high accuracy, several key factors contributed to this decision.

Initially, the LSTM autoencoder appeared to be a promising choice. This model, designed for sequential data like text, boasts LSTM memory cells that can effectively store and retrieve information. This ability to remember and forget information is invaluable in capturing anomalous patterns, making it well-suited for identifying unusual events. Moreover, the model excels at automatically learning and extracting meaningful features from text sequences, enabling it to detect complex patterns and deviations. However, as experimentation proceeded, limitations became apparent. The model struggled when processing long text sequences, performed sequential processing leading to slow training, and demanded significant computational resources. In the evaluation process, the LSTM Autoencoder showed promise for capturing anomalous patterns and extracting meaningful features from text sequences but faced limitations in handling long sequences, leading to the choice of the Google BERT model with its bidirectional context understanding and computational efficiency within the DFClaaS system. The Google BERT model, built on the transformer architecture, emerged as a strong contender. This model offers parallel processing and incorporates attention mechanisms for efficient text analysis. It is characterised by its bidirectional context understanding, extensive pre-trained knowledge, fine-tuning capabilities, and computational efficiency. However, it's worth mentioning that the specific model developed exhibited a slight increase in loss and a slight decrease in accuracy during validation.

Transitioning from the Google BERT model, another noteworthy approach worth recognizing is the gzip text classification algorithm, which achieved high accuracy by directly analysing text content and optimizing text representation using efficient data compression techniques.

The gzip text classification algorithm deserves recognition for its ability to achieve high accuracy by directly analysing text content, bypassing the need for complex neural network architectures. This approach leverages efficient data compression techniques to optimize text representation, allowing the classifier to work efficiently with compressed data. While the gzip approach showcases the potential for high accuracy, it is essential to acknowledge the significance of understanding language, context, and semantics, particularly in the context of text-based cybercrimes. Ultimately, the decision to employ the Google BERT model in the DFClaaS system aligns with the pursuit of a balance between accuracy, computational efficiency, and contextual understanding.

The subsequent section will bring our discussion to a close, summarizing the key findings and insights.

8. Conclusion

In the face of rapidly evolving digital threats, particularly those perpetrated through textual means, the DFClaaS system emerges as a beacon of hope. Taking a meticulous and multifaceted approach, it delved into diverse text-based cybercrime identification models, wielding the potent tools of Natural Language Processing (NLP) for both lexicon and sentiment analysis. This comprehensive expedition paved the way for a robust system capable of parsing the insidious language of cybercrime with high accuracy.

However, the journey, while fruitful, is far from over. As we forge ahead, refinement beckons the opportunity to improve the chosen model, optimizing its precision and minimizing potential shortcomings. In doing so, DFClaaS will cement its position as a crucial bridge between efficient language processing and robust cybersecurity, safeguarding users in the face of cybercrime's omnipresent shadow. By deciphering the intricate nuances of this clandestine language, DFClaaS offers invaluable insights into the dark underbelly of digital crime, illuminating a path toward a safer digital domain for all.

References

Stacey Omeleze Baror, Richard Adeyemi Ikuesan, and Hein S Venter. Functional architectural design of a digital forensic readiness cybercrime language as a service. In *ECCWS 2023 22nd European Conference on Cyber Warfare and Security*. Academic Conferences and publishing limited, 2023.

Rupa Ch, Thippa Reddy Gadekallu, Mustufa Haider Abidi, and Abdulrahman Al-Ahmari. Computational system to classify cyber crime offenses using machine learning. *Sustainability*, 12(10):4087, 2020.

Joel Chigada and Rujeko Madzinga. Cyberattacks and threats during covid-19: A systematic literature review. *South African Journal of Information Management*, 23(1):1–11, 2021.

KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17, 2011.

Malek Harbawi and Asaf Varol. The role of digital forensics in combating cybercrimes. In *2016 4th International Symposium on Digital Forensic and Security (ISDFS)*, pages 138–142. IEEE, 2016.

B Janet, Srinivasulu Reddy, et al. Anti-phishing system using lstm and cnn. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–5. IEEE, 2020.

Syed Mahbub and Eric Pardede. Using contextual features for online recruitment fraud detection. 2018.

Andiwi Meifilina, M Umanailo, and Imam Fachruddin. Cybercrime cases impact the development of communication technology that is troubling society. *SCOPUS–Q3*, 8(9):1224–1228, 2019.

Fabio Petroni, Tim Rocktaschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. ¨ Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Amber Stabek, Paul Watters, and Robert Layton. The seven scam types: mapping the terrain of cybercrime. In *2010 Second Cybercrime and Trustworthy Computing Workshop*, pages 41–51. IEEE, 2010.

Rana Mohtasham Aftab, Mariam Ijaz, Faisal Rehman, Ahmad Ashfaq, Hanan Sharif, Naveed Riaz, Shabbir Hussain, Muhammad Arslan, and Hadia Maqsood. A systematic review of the motivations of cyber-criminals and their attacking policies. In 2022 3rd International Conference on Innovations in Computer Science & Software Engineering (ICONICS), pages 1–6. IEEE, 2022.

Soumen Pal, Manojit Bhattacharya, Sang-Soo Lee, and Chiranjib Chakraborty. A domain-specific next-generation large language model (llm) or chatgpt is required for biomedical engineering and research. Annals of Biomedical Engineering, pages 1–4, 2023.

Jonah Berger and Grant Packard. Using natural language processing to understand people and culture. American Psychologist 77(4):525, 2022.

Thomas J Holt, Adam M Bossler, and Kathryn C Seigfried-Spellar. Cybercrime and digital forensics: An introduction. Routledge, 2022.

An Analysis of Cybersecurity Architectures

Noluntu Mpekoa
University of Johannesburg, South Africa

noluntum@uj.ac.za

Abstract: The 4th Industrial Revolution has increased high-capacity connectivity, new human-machine interactions largely with IoTs and smart devices. This digital revolution offers incredible conveniences such as the ability for users to access volumes of data, governments can address social challenges, connect remote villages in the country, and more. Once secluded systems are now connected and sharing information. This connectedness also poses some inconveniences as well, whenever a device joins the Internet, it becomes publicly discovered. Once these devices are discovered, they become open to cyberattacks. Cybersecurity has become a crucial part of daily life as cyberattacks have increased over time and have become more and more severe. The challenge that cybersecurity consultants find is the difficulty of measuring cybersecurity efforts in organizations. Another challenge could be finding a cybersecurity architecture that is effective and can fit different situations. The main aim of this study was to develop a comprehensive cybersecurity architecture that can be used by cybersecurity consultants when measuring cybersecurity effectiveness. This study conducted an in-depth literature review on current cybersecurity architectures offered by national and international cybersecurity organizations. The identified cybersecurity architectures that have been developed by other organizations were translated, interpreted, compared, and synthesized and a new cybersecurity architecture is proposed. The proposed cybersecurity architecture has the NIST goals as a foundation and the CIA triad at the center. The proposed cybersecurity architecture has domains such as application and Systems security, Information security, Network security, End-point security, Critical Infrastructure security, Mobile security, Storage security, etc. The proposed cybersecurity architecture seeks to assist cybersecurity consultants in answering questions from executives such as: Are we secure? Are security investments delivering value to the business? What is our preparedness for a cyberattack?

Keywords: Security architecture, Cybersecurity, Security capability domain, Cybersecurity control areas

1. Introduction

The Fourth Industrial Revolution (4IR) has led to an increase in high-capacity connectivity and new human-machine interactions, largely through the Internet of Things (IoT) and smart devices. This digital revolution offers incredible conveniences. Users can access vast amounts of data, governments can address social challenges, and remote villages can connect with the rest of the country. Communities have transitioned into digital spaces, encompassing work, leisure, and commerce (WEF, 2017; Ivanov & Das, 2020). Digital tools in the classroom have provided teachers and students with increased opportunities. Teachers have been pushed to evolve their teaching strategies to support personalized learning, creativity, innovation, and problem-solving, and provide more time for individual instruction. (WTO, 2020). Many businesses suffered during and after COVID-19, with some SMEs closing down. The 4IR has become more affordable and easier to implement, thus providing a viable recovery strategy for many businesses (Ivanov & Dolgui, 2021). PwC in 2018 conducted an annual global CEO survey, where "81% of executives agreed that 4IR technologies created new efficiencies, and 78% agreed that they helped them automate tasks" (PwC, 2018).

Previously isolated systems are now interconnected and sharing data. Möller (2023) states that this connectedness also poses some inconveniences as well, whenever a device joins the Internet, it becomes publicly discovered. Once these devices are discovered, they become open to cyberattacks (Singh & Kumar, 2020; Aphane, 2023). Cybersecurity has become an essential part of our daily lives due to the increasing frequency and severity of cyberattacks. Cybersecurity consultants face a significant challenge in measuring the effectiveness of cybersecurity measures in organizations. Another challenge could be finding a cybersecurity architecture that is effective and can fit different situations (Mbelli & Dwolatzky, 2016; Carcary, Doherty & Conway, 2019).

The main aim of this study was to develop a comprehensive cybersecurity architecture that can be used by cybersecurity consultants when measuring cybersecurity effectiveness. The following section will provide a brief overview of cybersecurity architectures, then followed by the research methodology utilised in this study. The proposed cybersecurity architecture is then presented and followed by the discussion and conclusion.

2. Overview of Cybersecurity Architectures

2.1 Cybersecurity Investment and ROI

The last quarter of 2022 and first quarter of 2023 saw an alarming 18.8% surge in cyberattacks in South Africa (SA). This is a clear indication of the ever-increasing threat of cybercrime and highlights the need for

individuals and organizations alike to take proactive measures to protect their digital assets (Kaspersky, 2023; DA, 2023). The Interpol (2022) report states that South Africa has been labeled as the cybercrime hub of Africa, and hackers on the dark web have shown interest in the country. In 2020, an Accenture report revealed that the country's internet users were less experienced and technically alert (Accenture, 2020). Poor cybersecurity awareness and lack of end-user training have contributed to successful cyberattacks in South Africa (Möller, 2023; Ngoma, Keevy, & Rama, 2021). Table 1 below, demonstrates the extent of imminent cyberattacks in the country, extracted from various sources (Paganini, 2022; DA, 2023)

Table 1: Summary of Cybersecurity Attacks in South Africa

YEAR	ATTACK DETAILS
March 2022	Everest gang demands $200K for data stolen from ESKOM
May 2022	TransUnion cyber-attack – hackers demand R225 million ransom
July 2022	Transnet Port Terminals attack
December 2022	Postbank loses over R18-million to cybercrime attacks
February 2023	Porsche South Africa suffers ransomware attack
April 2023	Justice department loses millions in yet another cyber attack
August 2023	Daily Maverick DDoS attack during BRICS Summit

Cybercrime is one of the biggest business risks in South Africa, and it is predicted that government departments will face an increasing number of attacks. The economy suffers an estimated loss of R2.2 billion per year due to cybercrime (Kaspersky, 2023; Paganini, 2022; Mbelli & Dwolatzky, 2016). Cyber security refers to the body of technologies, processes, and practices designed to protect networks, devices, programs, and data from attack, damage, or unauthorized access (Ivanov & Das, 2020; Singh & Kumar, 2020). Chief Information Officers (CIOs) and Chief Information Security Officers (CISOs) have experienced difficulties in justifying cybersecurity investments or expenses (Fowler & Chen, 2017; Hallman *et al.*, 2021; Moore, Dynes & Chang, 2016). Cybersecurity managers and specialists are constantly asked questions by executives such as: Are we secure? Are security investments delivering value to the business? What is our preparedness for a cyberattack? The primary responsibility of CIOs/CISOs is to implement effective cybersecurity measures to prevent attacks that may disrupt or disable the operations of their organization.

2.2 Cybersecurity Architectures

Cybersecurity architectures consist of distinct and comprehensive domains that are separate from each other. The different categories that come under cybersecurity are often referred to as domains. Some people may also call them control areas or capability domains (Roy, Kim & Trivedi, 2010; Möller, 2023). When referring to cybersecurity domains, one is essentially talking about the specific area of cybersecurity being discussed. In simpler terms, a domain refers to a specific category or focus area, especially in cybersecurity. When various cybersecurity domains are integrated, it results in the formation of a Cybersecurity Architecture. Cybersecurity organizations and frameworks define the domains of cybersecurity differently (Möller, 2023, Bokan & Santos, 2021; AlHamdani, 2020). The concept of cybersecurity architecture has many faces, and each framework has its own focus and strengths. To such an extent that, several frameworks exist for cybersecurity architecture, the most popular and common ones are the NIST Framework, Open-Source Architecture and the International Information System Security Certification Consortium, otherwise known as (ISC)2.

Some of these cybersecurity architectures complement each other and may have overlapping features. Application security, physical security, risk assessment, and threat intelligence are some of the most common domains in cybersecurity. Table 2 below lists some of the common cybersecurity architectures.

Table 2: Cybersecurity Domains

NIST Capability Domains	CISSP Certification	GSEC Certification	Open-Source Architecture	Decipher Zone
Access Control	Security and Risk Management	Active Defense	IT Service Security Patterns	Network Security
Security Assessment	Asset Security	Cryptography	IT Application-Level patterns	Cloud Security

NIST Capability Domains	CISSP Certification	GSEC Certification	Open-Source Architecture	Decipher Zone
Awareness & Training	Security Engineering	Defensible Network Architecture	IT Infrastructure Patterns	Application Security
Risk Management	Communications and Network Security	Incident Handling and Response	Central Security Services	Critical Infrastructure Security
Situational Awareness	Identity and Access Management	Linux Security	Governance	End-User Security
Asset Management	Security Assessment and Testing	Security Policy		Disaster Recovery
Media Protection	Security Operations	Web Communication Security		Information security
Maintenance	Software Development Security			Storage Security
Personal Security				Mobile Security
Audit & Accountability				
System & communication protection				
Identification & Authentication				

The task at hand is to design a cybersecurity framework that is both effective and tailored to one's unique situation or organization. The second challenge that Cybersecurity managers and specialists face is creating the best cybersecurity architecture from a diverse range of options.

3. Research Methodology

The main aim of this study was to develop a comprehensive cybersecurity architecture that can be used by cybersecurity consultants when measuring cybersecurity effectiveness. The proposed cybersecurity architecture seeks to assist cybersecurity consultants in answering questions from executives such as: Are we secure? Are security investments delivering value to the business? What is our preparedness for a cyberattack? An in-depth literature study was conducted to have a fundamental understanding of the research problem. A qualitative meta-synthesis methodology to synthesize different cybersecurity architectures was utilised. This process follows the steps in the meta-ethnography approach (Bhattacharyya, 2009; Rajasekar & Verma, 2013; Snyder, 2019). Qualitative meta-synthesis is a technique for synthesizing the findings of qualitative research. The method aims to explain the findings of similar qualitative studies. The result is used as the basis for the design of the proposed comprehensive cybersecurity architecture (Goddard & Melville, 2004). The primary sources of data were derived based on published articles, reviews, and books to gain essential data that assisted in formulating a suitable cybersecurity architecture. After a comprehensive and iterative literature search, several cybersecurity architectures were identified (Table 2). Most of the cybersecurity architectures were developed by cybersecurity organizations based on reports and white papers from consulting firms and international organizations. The cybersecurity architectures identified have been published within the last ten years, from 2013 to 2023, and are based on different perspectives.

4. Proposed Cybersecurity Architecture

Cybersecurity solutions should address each step with a targeted solution while maintaining an overall focus on security. Having a general sense of cybersecurity framework is no longer enough. An intricate network of security measures, layered atop one another, consolidates cybersecurity domains. Cybersecurity domains is a term used to describe different areas where cybersecurity practices can be applied. Each aspect of the cyber domain presents unique security challenges and risks that require attention. To ensure cybersecurity,

organizations must identify and mitigate risks associated with every subdomain. This holistic view of organizational information security allows cybersecurity specialists to apply different approaches to different parts of cyber domains.

4.1 Proposed Cybersecurity Architectures

A coordinated effort is vital in maintaining a good cyber security posture as organizational assets are comprised of various platforms. The web of cybersecurity measures working together to protect a system is called cybersecurity domains.

The proposed cybersecurity architecture makes use of the NIST framework (Identify, Protect, Detect, Respond, Recover) "GOALS" as the foundation (these are further discussed in Section 5.1) and utilises the pillars from the CIA triad (Confidentiality, Integrity, Authenticity) at the CENTER (discussed in Section 5.2) and finally offer nine (9) subdomains as the core of the architecture (discussed in Section 5.3).

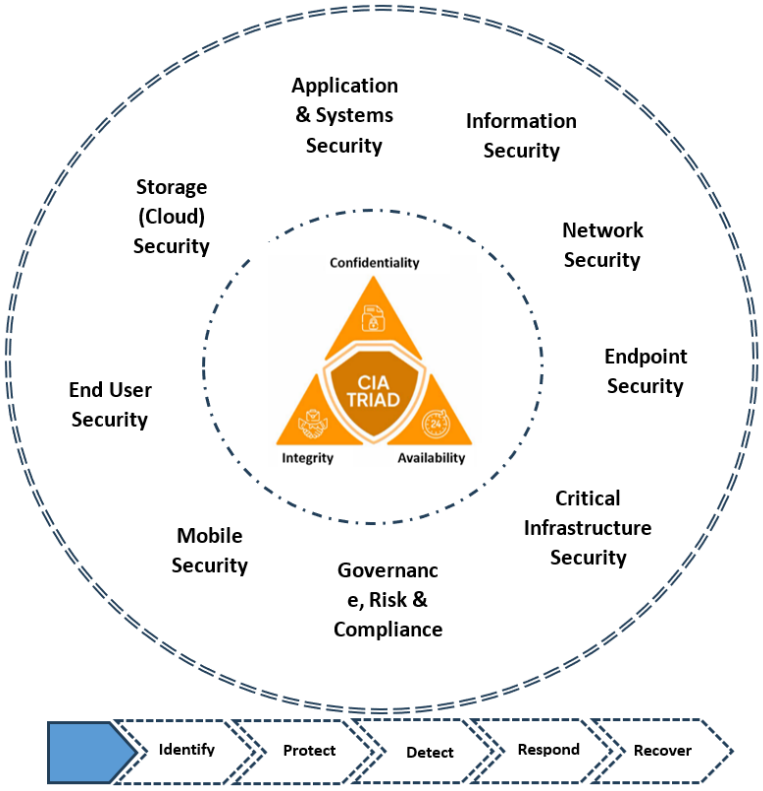


Figure 1: Proposed Comprehensive Cybersecurity Architecture

The proposed cybersecurity architecture can be refined over time to ensure it is comprehensive and covers every angle /inch of an organization.

5. Discussion on Proposed Cybersecurity Architecture

This section discusses the proposed cybersecurity architecture, motivating each component making the architecture. This section discusses the five NIST functions, followed by the CIA Triad, and finally the nine (9) cybersecurity domains.

5.1 NIST Goals

The National Institute of Standards and Technology (NIST) is a worldwide recognised information security best practice framework (Almuhammadi & Alsaleh, 2017). The NIST Information Security Framework consists of a collection of industry standards and best practices. The NIST framework comprises of five functions that operate concurrently and continuously (Möller, 2023; Malatji, Marnewick & Von Solms, 2022). When implemented together, these functions provide a comprehensive and strategic view of the cybersecurity risk management program (Curtis & Mehravari, 2015, Le & Hoang, 2017). The NIST information security framework offers the following five (5) different functions/goals:

- IDENTIFY**:** refers to tools to track and trace every IT asset running within an organization (Almuhammadi & Alsaleh, 2017)
- PROTECT**:** refers to tools used to protect all identified assets (Sulistyowati, Handayani & Suryanto, 2020)
- DETECT**:** refers to tools for monitoring the organisation and detecting intruders (Möller, 2023)
- RESPOND**:** refers to automated tools that flag incidents and direct ticket assignments accordingly (Le & Hoang, 2017)
- RECOVER**:** refers to rapid recovery to normal functionality after a cyberattack (Curtis & Mehravari, 2015)

5.2 CIA Triad

Also known as the pillars of information security, the CIA triad comprises of Confidentiality, Integrity and Availability. The terms 'confidentiality', 'integrity' and 'availability' have been widely used in the information security practice and in academic literature to help evaluate and create information systems (Qadir & Quadri , 2016; Fenrich, 2008). CIA triad still continues to assume a major role in information security practice as they offer the fundamental elements of security controls in information systems (Samonas & Coss, 2014). Essentially, this means that a sound information security system must protect confidentiality, integrity, and availability. The CIA triad has not only shaped and informed the theoretical understanding of information security, but also the very practices through which security is developed and implemented in organizations (Hiza, 2022). Following are definitions of each pillar of information security:

- CONFIDENTIALITY**:** this refers to the protection of information from unauthorized or unintended disclosure (Lundgre n & Möller, 2019; Fenrich, 2008)
- INTEGRITY**:** this refers to the prevention of unauthorized information alteration or destruction (Hiza, 2022; Samonas & Coss, 2014)
- AVAILABILITY**:** this refers to judicious and reliable access and use of data and information, whenever and wherever it is needed by the appropriate parties (Qadir & Quadri, 2016; Hiza, 2022)

5.3 Cybersecurity Domains

The last component of the proposed cybersecurity architecture are the nine (9) cybersecurity domains. Each domain should satisfy confidentiality, integrity and availability, as well as the five NIST goals (identify, protect, detect, respond and recover). The following section discusses the nine cybersecurity domains:

APPLICATION AND SYSTEM SECURITY**:** involves creating, adding, and testing security features in applications to prevent cyber-attacks and make them secure. Ensures security of existing software and applications and defends against threats during the application's development stage. This allows organisations to install many layers of security within a software and to any information related to it, also limit unwanted changes to the software information (Usman, Jan, He & Chen, 2019; Sabillon, Serra-Ruiz, Cavaller and Cano, 2017).

INFORMATION SECURITY**:** this domain protects all kinds of data from unauthorized access and prevents any changes, modifications, disclosures, and deletions. It protects sensitive business information and ensures data security (Obrst, Chase & Markeloff, 2012).

NETWORK SECURITY**:** this involves protecting computer networks and network-accessible resources from unauthorized access and modifications. Network security combines strategies, technologies, and processes that protect a network from intrusion (Sabillon, Serra-Ruiz, Cavaller and Cano, 2017; Usman, Jan, He & Chen, 2019).

ENDPOINT SECURITY: the protection of end-user devices such as desktops, laptops, mobile phones, and tablets from malicious threats and cyberattacks. This includes scanning all the devices from a network for threats, anomalies, and suspicious behaviour (Wang, Lau & Gerdes, 2018; Obrst, Chase & Markeloff, 2012).

CRITICAL INFRASTRUCTURE SECURITY**:** refers to protecting the infrastructure without which the security of a nation may be in danger. It consists of both cyber and physical systems crucial to ensuring cyber security. The cybersecurity of any organization can run smoothly when the physical assets of the company are safeguarded (Collier, Linkov & Lambert, 2013).

GOVERNANCE, RISK & COMPLIANCE**:** refers to the development and creation of company policies on various aspects of internal and external security. The strategy of any organization to deal with potential and real cyber

security attacks. This includes the development of preventive and recovery systems to manage and respond to potential cyber threats (Usman, Jan, He & Chen, 2019; Obrst, Chase & Markeloff, 2012).

MOBILE SECURITY**:** protects mobile devices or tablets from any malicious threats of data loss or asset loss. With the popularity of mobile, it has become an essential part of cybersecurity (Sabillon, Serra-Ruiz, Cavaller and Cano, 2017).

END-USER SECURITY**:** training employees and people on cybercrimes, and industry best practices and protecting themselves from threats like social engineering are incredibly important. This will either prevent unnecessary attacks or enable individuals to identify risks themselves (Wang, Lau & Gerdes, 2018).

CLOUD SECURITY**:** protects the cloud infrastructure, environment, applications, and data from threats. It ensures the authentication of end users and devices and provides data privacy protection (Wang, Lau & Gerdes, 2018; Obrst, Chase & Markeloff, 2012).

5.4 Benefits of Proposed Cybersecurity Architecture

Security professionals must present cybersecurity data in a way that satisfies executives' curiosity, leads to actionable insights, and is easy to understand. The benefits of the proposed cybersecurity architecture include:

- *A comprehensive cybersecurity structure:* the proposed cybersecurity architecture is an amalgamation of numerous cybersecurity frameworks merging their strengths and benefits.
- *Offers a bird's eye view of the organisation:* the proposed cybersecurity architecture allows one to work on multiple plains of cyberattacks, all at once.
- *Offers better control over the environment:* the proposed cybersecurity architecture allows one to determine risk tolerance and set controls.
- *Contextualise cybersecurity solution:* Depending on the context and requirements of an environment, domains can be added or removed from the proposed cybersecurity architecture. However, some organizations are more advanced than others.

Security professionals need to carefully choose the most appropriate metrics while presenting reports to business teams. It is important to report the outcomes of cybersecurity solutions and their impact on organization's assets, operations, and reputation. A combination of quantitative and qualitative measures can be used, such as key performance indicators (KPIs), key risk indicators (KRIs), return on investment (ROI), cost of risk (COR), customer satisfaction, and stakeholder feedback.

The ability to quickly learn from failures, adapt and repurpose is critical to successfully measure cybersecurity.

6. Conclusion

The recent high-profile security breaches of organizations have demonstrated the need for good cybersecurity measures. These breaches have resulted in the loss of incredibly sensitive user information, causing irreversible financial and reputational damage. Every day, attackers target companies of all sizes to steal sensitive information or disrupt operations. Companies are often unaware of the many threats that exist within their IT infrastructure, and as a result, they do not implement cyber security countermeasures until it is much too late. In the near future, it is unlikely that the frequency or intensity of cyberattacks will decrease. Attacks on critical infrastructure are expected to increase in both strength and sophistication over time. The aim of the study was twofold: first, it provides an in-depth review of cybersecurity architectures; second, it proposes a cybersecurity architecture that extends existing cybersecurity architectures. The in-depth review features five most popular and widely used cybersecurity frameworks. The proposed cybersecurity architecture has the NIST goals as a foundation and the CIA triad at the center. The proposed cybersecurity architecture has domains such as application and Systems security, Information security, Network security, End-point security, Critical Infrastructure security, Mobile security, Storage security etc. The proposed cybersecurity architecture seeks to assist cybersecurity consultants in answering questions from executives such as: Are we secure? Are security investments delivering value to the business? What is our preparedness for a cyberattack? The benefits of the proposed cybersecurity architecture include a comprehensive cybersecurity architecture; offers a bird's eye view of the organisations' security posture; offer better control of the environment and an architecture that can be tailored to meet organisational needs and requirements.

References

Almuhammadi, S., and Alsaleh, M. (2017). Information security maturity model for NIST cyber security framework. Computer Science & Information Technology (CS & IT), 7(3), 51-62.

Aphane, M. P. (2023). Cybersecurity Awareness on Cybercrime Among the Youth in Gauteng Province. International Journal of Social Science Research and Review, 6(8), 23-32.

Bhattacharyya, D. K. (2009). Research methodology. Excel Books India.

Carcary, M., Doherty, E., and Conway, G. (2019). A framework for managing cybersecurity effectiveness in the digital context. In European Conference on Cyber Warfare and Security (pp. 78-86). Academic Conferences International Limited.

Collier, Z. A., Linkov, I., and Lambert, J. H. (2013). Four domains of cybersecurity: a risk-based systems approach to cyber decisions. Environment Systems and Decisions, 33, 469-470.

Curtis, P. D., and Mehravari, N. (2015). Evaluating and improving cybersecurity capabilities of the energy critical infrastructure. In 2015 IEEE international symposium on technologies for homeland security (hst) (pp. 1-6). IEEE.

Fenrich, K. (2008). Securing your control system: the" CIA triad" is a widely used benchmark for evaluating information system security effectiveness. Power Engineering, 112(2), 44-49.

Fowler, S., and Chen, P. P. (2017). CsPI: A New Way to Evaluate Cybersecurity Investments: A Position Paper. In 2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C) (pp. 283-284). IEEE.

Goddard, W., and Melville, S. (2004). Research methodology: An introduction. Juta and Company Ltd.

Hallman, R. A., Major, M., Romero-Mariona, J., Phipps, R., Romero, E., Slayback, S. M., and San Miguel, J. M. (2021). Determining a return on investment for cybersecurity technologies in networked critical infrastructures. International Journal of Organizational and Collective Intelligence (IJOCI), 11(2), 91-112.

Hallman, R. A., Major, M., Romero-Mariona, J., Phipps, R., Romero, E., John, M., and Miguel, S. (2020). Return on Cybersecurity Investment in Operational Technology Systems: Quantifying the Value That Cybersecurity Technologies Provide after Integration. In COMPLEXIS (pp. 43-52).

HIZA, D. (2022). Assessing the Significance of Cia Triad Security Model in Establishing ICT Security Controls in The Public Sector (Doctoral dissertation, Institute of Accountancy Arusha (IAA)).

Ivanov, D. and Das, A. (2020). 'Coronavirus (COVID-19/SARS-CoV-2) and supply chain resilience: A research note', International Journal of Integrated Supply Management 13(1), 90–102. https://doi.org/10.1504/IJISM.2020.107780

Ivanov, D. and Dolgui, A. (2021). 'A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0', Production Planning & Control 32(9), 775–788. https://doi.org/10.1080/09537287.2020.1768450

Le, N. T., and Hoang, D. B. (2017). Capability maturity model and metrics framework for cyber cloud security. Scalable Computing.

Lundgren, B., and Möller, N. (2019). Defining information security. Science and engineering ethics, 25, 419-441.

Malatji, M., Marnewick, A. L., and Von Solms, S. (2022). Cybersecurity capabilities for critical infrastructure resilience. Information & Computer Security, 30(2), 255-279.

Mbelli, T. M., and Dwolatzky, B. (2016). Cyber security, a threat to cyber banking in South Africa: An approach to network and application security. In 2016 IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud) (pp. 1-6). IEEE.

Möller, D. P. (2023). NIST Cybersecurity Framework and MITRE Cybersecurity Criteria. In Guide to Cybersecurity in Digital Transformation: Trends, Methods, Technologies, Applications and Best Practices (pp. 231-271). Cham: Springer Nature Switzerland.

Moore, T., Dynes, S., and Chang, F. R. (2016). Identifying how firms manage cybersecurity investment. In Workshop on the Economics of Information Security (WEIS) (pp. 1-27).

Obrst, L., Chase, P., and Markeloff, R. (2012). Developing an Ontology of the Cyber Security Domain. In STIDS (pp. 49-56).

PwC (2018). Navigating the rising tide of uncertainty, 23rd Annual Global CEO Survey. Accessed 09 November 2023 at https://www.pwc.com/gx/en/ceo-agenda/ceosurvey/2020.html

Qadir, S., and Quadri, S. M. K. (2016). Information availability: An insight into the most important attribute of information security. Journal of Information Security, 7(3), 185-194.

Rajasekar, D., and Verma, R. (2013). Research methodology. Archers & Elevators Publishing House.

Roy, A., Kim, D. S., and Trivedi, K. S. (2010). Cyber security analysis using attack countermeasure trees. In Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research (pp. 1-4).

Sabillon, R., Serra-Ruiz, J., Cavaller, V., and Cano, J. (2017). A comprehensive cybersecurity audit model to improve cybersecurity assurance: The cybersecurity audit model (CSAM). In 2017 International Conference on Information Systems and Computer Science (INCISCOS) (pp. 253-259). IEEE.

Samonas, S., and Coss, D. (2014). The CIA strikes back: Redefining confidentiality, integrity and availability in security. Journal of Information System Security, 10(3).

Singh, S., and Kumar, S. (2020). The times of cyber attacks. Acta Technica Corviniensis-Bulletin of Engineering, 13(3), 133-137.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of business research, 104, 333-339.

Sulistyowati, D., Handayani, F., and Suryanto, Y. (2020). Comparative analysis and design of cybersecurity maturity assessment methodology using nist csf, cobit, iso/iec 27002 and pci dss. JOIV: International Journal on Informatics Visualization, 4(4), 225-230.

Usman, M., Jan, M. A., He, X., and Chen, J. (2019). A survey on representation learning efforts in cybersecurity domain. ACM Computing Surveys (CSUR), 52(6), 1-28.

Wang, H., Lau, N., and Gerdes, R. M. (2018). Examining cybersecurity of cyberphysical systems for critical infrastructures through work domain analysis. Human factors, 60(5), 699-718.

World Economic Forum (WEF). (2017). Impact of the Fourth Industrial Revolution on supply chains, System Initiative on Shaping the Future of Production, World Economic Forum, Geneva.

World Trade Organisation (WTO). (2020). E-commerce, trade and the Covid-19 pandemic, pp. 1–8, WTO Secretariat, Geneva.

Bibliometric Analysis of Cyber Warfare Research in Africa: Landscape and Trends

Jabu Mtsweni[1,2] and Mphahlela Thaba[1]
[1]Council for Scientific and Industrial Research, Pretoria, South Africa
[2]Stellenbosch University, Security Institute for Governance and Leadership in Africa, Faculty of Military Science, Saldanha, South Africa

mtswenij@gmail.com
jthaba@csir.co.za

Abstract: As the digital landscape continues to evolve, cyber warfare has emerged as a prominent domain of warfare, with superpower nations actively demonstrating their capabilities in the cyberspace. This study posits that African countries exhibit a relative lag in research and development of cyber warfare capabilities, as evidenced by the absence of African nations in the National Cyber Power Index released by the Belfer Centre for Science and International Affairs in 2022. To address this knowledge gap, this paper presents a comprehensive bibliometric analysis of cyber warfare research and development within the African continent. The analysis aims to illuminate research productivity, performance, science mapping, and key contributors at both national and institutional levels. It seeks to uncover thematic trends, pinpoint key research areas, and identify research connections within the African context. This research evaluates the African continent's research participation and development in the cyber and/or information warfare domain over the past 23 years. The analysis encompasses scholarly articles and conference proceedings published between 2000 and 2023, utilizing Scopus as the primary data source. Preliminary findings suggest that cyber warfare research in Africa is concentrated in a limited number of countries, with South Africa emerging as the leading contributor. A comparative analysis further reveals that developed countries generally outpace African nations in cyber warfare research and development, corroborating the rankings presented in the National Cyber Power Index (NCPI) and Global Cybersecurity Index (GCI).

Keywords: Cyber security, Cyber warfare, Cyber operations, Cyber defence, Cyber attacks, Information warfare

1. Introduction

African nations face significant challenges in establishing and operationalizing cyber warfare capabilities (Thaba & Mtsweni, 2023). The Global Cybersecurity Index (GCI), a benchmark assessment of countries' cybersecurity commitments conducted by the International Telecommunications Union (ITU), highlights the relative underdevelopment of cybersecurity in many African countries (ITU, 2021). This underdevelopment manifests in several areas, including: (1) inadequate legal frameworks to support cybersecurity efforts; (2) insufficient technical measures to safeguard against cyberattacks and threats; (3) underdeveloped organizational structures for managing cyber risks and responding to incidents; (4) limited capacity building initiatives to cultivate a skilled cybersecurity workforce; and (5) weak international cooperation mechanisms for sharing cyber threat intelligence and conducting joint cybersecurity exercises.

In contrast, the National Cyber Power Index (NCPI), which focuses on the offensive aspects of cybersecurity and includes at least 30 countries, features only one African country (i.e., Egypt) (Voo, et al., 2022). This disparity reflects the dearth of research and development in cyber warfare capabilities within Africa, compared to countries like the United States, China, and Russia where cyber warfare capabilities are advanced and operationalized.

This research employs a bibliometric analysis technique to advance the comprehension of cyber warfare research in Africa, offering a valuable resource for scholars and policymakers alike. Through a comprehensive assessment of research productivity, performance, and regional dynamics, this study lays the groundwork for fostering collaboration, innovation, and knowledge dissemination within Africa's evolving cyber warfare landscape.

The rest of this paper is structured as follows: Section 2 provides context and overview on the definitions of terms used in this paper as well as the literature review. In Section 3, the research methodology adopted is presented with the bibliometrics analysis process explained, including data collection profiles. Section 4 presents the results of the bibliometric analysis using VOSviewer and Bibliometrix tool. In Section 5, a high-level comparative analysis is presented to put the bibliometric analysis of African countries on cyber warfare research

in perspective. Section 6, discusses the research results, limitations, and implications. The paper is concluded with a summary and future research recommendations in Section 7.

2. Cyber Warfare: Literature Review

As explained by Thaba and Mtsweni (2023), the cybersecurity domain is quite broad and encompasses various interpretations and definitions depending on the context. Therefore, providing a clear definition and context for cyber warfare is crucial.

2.1 Definitions and Context

In this paper, we define *cyber warfare* as the defensive and offensive capabilities of a nation-state to influence, protect, manipulate, or exploit cyberspace for strategic national objectives. Moreover, the terminologies used in cyber warfare research vary across studies (Lukin, 2019; Colarik, 2017). For instance, cyber warfare is sometimes spelled as cyberwar (one word), which can impact bibliometric analysis. Additionally, in some countries, the term "cyber defense" is written with a hyphen ("cyber-defense"), while in others, it is written without a hyphen ("cyber defense"), and in other cases with the UK spelling (“cyber defence”). These variations are quite common in the cybersecurity terminologies. In this research, we attempted to consider these different variations in our data collection. Nevertheless, we accept that it will remain challenging to catch-all the different variations within the search queries.

2.2 Cyber Warfare Research in the 21st Century

The development of cyber warfare capabilities by nation states is evident even though African nations seem to be lagging. The modern conflicts observed in Europe, Asia, Africa, and other continents are showing that kinetic warfighting is increasingly supported by cyber operations at different levels. The Russia-Ukraine war is a typical example where different cyber dimensions have been demonstrated (Willet, 2023). The conflict between Israel and Palestine has seen cyber-attacks escalate between nations, including a shutdown of Iran’s gas stations by the Israeli hacker group (The Times of Israel, 2023). In Europe, the North Atlantic Treaty Organization (NATO) agreed in 2016 to recognize cyber warfare as an operational domain (Binnendijk, et al., 2019), and this is a clear demonstration that cyber warfare will continue to play a role in modern-conflicts and African countries should not be found wanting.

It is also evident that some countries, through foreign nation-state actors are already using cyber-weapons as means to entrench their economic influence and/or governments’ manipulation in Africa. Several incidents have been observed in Mozambique (Thaba & Mtsweni, 2023), Kenya (Fielder, 2021), Nigeria, Ethiopia, South Africa, Sudan, and Ghana in the recent past (Amoah, 2019; Mtsweni & Thaba, 2021).

Over and above the common cyber defence capabilities, developed nations have been building offensive capabilities (Voo, et al., 2022) arranged through national structures such as Cyber Command. These capabilities include: (1) Cyber surveillance and monitoring, (2) Cyber foreign intelligence gathering, (3) Information control and manipulation, (4) systems destruction and hacking, and (5) cyber competence development, amongst others.

2.3 Related Work

Bibliometric analysis (Donthu, et al., 2021) studies have been exponentially increasing across all research disciplines since 2019, with close to 50,000 research results on Google Scholar between 2019-2023 (4 years), compared to approximately 41,100 between 2010-2018 (8 years). This suggests that bibliometric analysis studies have taken off because of their nature of providing insights and trends. Google trends between 201-2023 also confirms this exponential growth (see Figure 1).

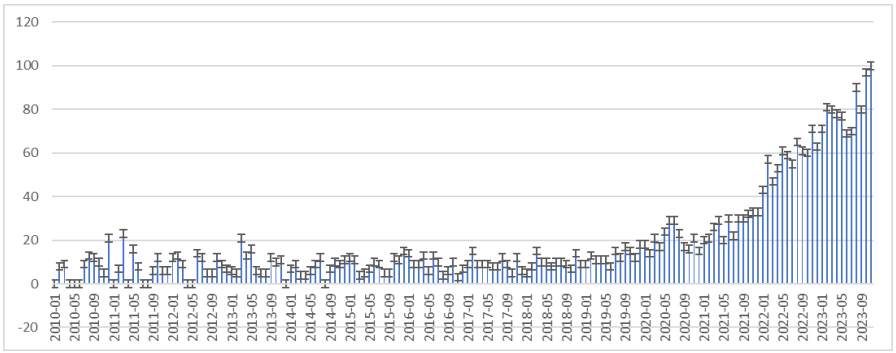


Figure 1: Google Trends on bibliometric analysis (2010-2023) (x: interest in % and y-years)

In the realm of cybersecurity, there is a growing trend in employing bibliometric analysis in research studies. Rahim (2021) utilized this technique to investigate the trends in literature regarding cyber threats and cyber-attacks within the Higher Education context from 2004 to 2019 on a global scale. The study encompassed a comprehensive analysis of scientific performance in this domain, revealing that South Africa ranked at #14, while other African nations, such as Nigeria and Morocco, find themselves towards the bottom of the top 100 in relation to cyber threats and attacks research publications.

Nobanee, et al., (2023) conducted a bibliometric analysis on literature related to cybercrime and cybersecurity risks, evaluating global research developments. This comprehensive study analysed 749 articles spanning from 1999 to 2021, utilizing the Scopus database and VOSviewer platform. The findings of the study showed a growing trend on cybersecurity risks even though some research gaps still exist in the field.

Kyrdoda, et al., (2023) delved into cybersecurity technology research from 2011 to 2021 using bibliometric analysis, discovering a notable surge in interest with over 1,100 papers published in the past decade. Interestingly, the analysis demonstrated the absence of African authors in the top contributors to cybersecurity research, with no African country ranking in the top 15 publications, each boasting a citation count of over 42 per document. The United States and China emerged as leaders in cyber technology research during the specified timeframe.

Tse, et al., (2015) employed the bibliometric analysis technique to explore trends and patterns in privacy and ethics within the IEEE Security and Privacy journal, focusing on the period from 2003 to 2014. While shedding light on certain aspects of cyber warfare, the study did not unveil any insights related to the African continent.

Through the literature review, it is apparent that no bibliometric analysis has been conducted on cyber warfare research specifically in Africa, nor on general cybersecurity trends. Thus, the current research study is deemed relevant and timely, aiming to fill this gap by providing valuable insights into the research and development trends within African countries.

3. Research Methodology

In this paper, we employed a bibliometric analysis technique to study cyber warfare research, developments, and capabilities in Africa spanning the past 23 years (2000-2023). The choice of bibliometric analysis as the research methodology stems from its widespread use in business, academia, and various domains to understand trends over time.

Bibliometric analysis is defined as a quantitative method for gauging the research impact and productivity of scholars and institutions (Donthu, et al., 2021). It entails analysing citations in scholarly publications to establish connections to other works or researchers and is widely applied in fields such as library and information science and research evaluation (van Eck & Waltman, 2010; Aria & Cuccurullo, 2017).

The benefits of bibliometric analysis are manifold. Its quantitative nature allows for objective measurements, and its transparency, utilizing publicly available data, facilitates easy replication of results within the same domain and period. Furthermore, its scalability enables the analysis of numerous publications over an extended period (Donthu, et al., 2021).

However, it is crucial to note the limitations of bibliometric analysis, particularly in cases where research data or information about a topic is not published in scientific publications or publications are not indexed by databases

such as Scopus or Web of Science. Consequently, this study was restricted to the research and development of published cyber warfare research data indexed by Scopus.

To complement the bibliometric analysis, we enhanced the research findings through a comparative analysis (Blair-Walcott, 2023) against the top five cyber power countries based on NCPI for research validation.

3.1 Stages of Bibliometric Analysis

In conducting the bibliometric analysis, a standard process was followed as depicted in Figure 1 below. This process was adapted from (Donthu, et al., 2021).



Figure 2: Bibliometric Analysis Stages (source: authors)

- Conducting a bibliometric analysis involves a systematic process of searching, analyzing, and interpreting data to identify trends, patterns, and insights in scholarly literature. The following are the steps followed.
- **Step 1: Define research question and objectives for bibliometric analysis:** for the research study, we defined the research question as "***what are the trends in research on cyber warfare in Africa over the past 23 years?***". The objective for this research was to identify and understand the cyber warfare research capability development trends and landscape in Africa in the context of globalization and digital transformation.
- **Step 2: Choose techniques and databases for bibliometric analysis:** we chose performance analysis, science mapping, and network analysis techniques as appropriate for this research study. In performance analysis, our focus was on publication, citation, and collaboration analyses. In science mapping, we focused on authors' relationships, topics, and thematic analyses. We also opted for network analysis by conducting clustering and visualization on the collected data as demonstrated in Section 4. We opted to using Scopus for data collection. This was because at the time of this research, it was able to support an export of a large-data set (up to 20,000 records at a time) compared to Web of Science. It is also one of the most trusted sources of research data and impact analysis.
- **Step 3: Formulate search query and collect the bibliometric data**: the keywords defined for the search query were: *"Cyber War*" OR "Cyber Defen*" OR "Cyber Offen*" OR " Cyber Oper*" OR "Cyberwar*" OR "Infowar*" OR "Information War*" OR "Cyber-war*" OR "Cyber-Defen*" OR "Cyber-Offen*".* There are still several challenges in the cyber domain in terms of agreed-upon taxonomy, terminologies, and definitions and for our search query, we attempted to be as inclusive as practical on the different variations in the cyber warfare domain. We used Boolean operator "OR" in Scopus to combine keywords and phrases closely related to the research topic. Data collection focused on final and in the press publications in conference papers, research articles, book chapters, editorials, letters, and surveys.
- **Step 4: Collect and refine the bibliometric results**: we collected the data and reviewed the search results. We then refined the data through the identification of relevant publications, countries, and affiliations for further analysis. We finally extracted the sampled data through a comma-separated values (CSV) file.
- **Step 5: Analyze, interpret, and visualize the bibliometric analysis**: the extracted CSV data file was imported into the VOSviewer platform (van Eck & Waltman, 2010) for visual analysis, Bibliometric tool (Aria & Cuccurullo, 2017) for other analysis not found in VOSviewer. Excel Data Analysis (Microsoft, 2021) was also employed in cases where we noticed duplicates and for cleaning and merging of data with same names that had different variations.

3.2 Bibliometric Analysis Tools

Biometric analysis is a complex process since it involves several technical steps and large datasets. Over the years, researchers have, as a result, developed open-source tools that simplifies and support the analysis process. In this study, we benchmarked a series of tools for this purpose as summarized below.

- **VOSviewer** (van Eck & Waltman, 2010) is a free software platform for constructing and visualizing bibliometric networks. It can be used to create maps of co-authorship, co-citation, and co-occurrence networks. VOSviewer is particularly well-suited for analysing large datasets, and it can produce high-quality visualizations that are easy to interpret.
- **CiteSpace** (Chen, 2014) is a free Java-supported tool for visualizing and analysing trends and patterns in scientific literature. It is primarily used for Web of Science data, but it can also import data from other sources such as arXiv and PubMed. CiteSpace is particularly useful for identifying emerging trends and hotspots in research.
- **Bibliometrix** (Aria & Cuccurullo, 2017) is a versatile R-tool package for performing comprehensive science mapping analysis of scientific literature suitable for handling large datasets. It provides a wide range of command-line and web-based tools for data import, cleaning, and analysis, as well as for creating visualizations.
- **Science of Science (Sci2)** (Lewis & Alpi, 2017) is a modular and open-source toolset specifically designed for the study of science. It includes a wide range of tools for data collection, analysis, and visualization. Sci2 is particularly well-suited for analysing large-scale scientific datasets.

For this study, we opted for the VOSviewer platform and Bibliometrix R-tool because these tools were free to use, easy to access and setup, demonstrated wide usage in literature, and provided for different visualizations.

4. Bibliometrics Analysis Results

As of 26 November 2023, an initial total of 33,420 research documents were retrieved using the Scopus database search engine as per the defined search keywords. The date range used was between 2000-2023. The search scope was within “all the fields” as defined by Scopus. When the search scope was limited to “article title, abstract, and keywords” only 6,091 documents were found, globally. A total of 4,727 documents were found based on the search query on the “abstracts” of the research publication. Limiting the search results to only the relevant keywords found in the research articles, a total of 2,774 documents were extracted from different research institutions across the globe. When we limited the search query to the relevant “article title”, 1,878 articles were found.

Employing different layers of filtering on search queries is crucial for an objective understanding of cyber warfare research trends in Africa, preventing overstating or understating the search results. Thoroughness is also essential to avoid misinterpreting the results due to the limitations of keyword searches. Using only one type of search query can lead to duplicated or irrelevant results. For example, searching within “all fields" in Scopus articles may yield results where the defined keywords appear in sections unrelated to cyber warfare. Conversely, using the "article title" search may exclude relevant publications if they do not include the keywords in the title.

For purposes of the analysis, we opted to use bibliometric data for Africa retrieved within a combination of “Article title, Abstract, and Keywords” search query. This, we believe may provide a general overview and insights of the cyber warfare research trends in Africa over the past 23 years.

We subsequently excluded retracted publications, raw data papers, and erratum publications from the analysis. The data was restricted English-written publications; however, this language limitation was not considered for the comparative analysis. Furthermore, we filtered the results by focusing on the identifiable African countries.

A total of 154 documents were discovered using this limited search query purely focusing on cyber warfare related research in Africa. From the analysis, it appears that the first indexed publication on Scopus on cyber warfare for Africa is from 2005 as the data returned contained years 2005-2023. This means our analysis covered 18 years instead of the 23 years as per the research objectives.

4.1 Publication Analysis

Focusing on the African content, only 154 publications written by over 254 co-authors were found on Scopus focusing on the subject matter. An estimated 68% of the research in cyber warfare originates from South Africa and the rest from other African countries.

Most publications on cyber warfare research in Africa are published in conference proceedings (61.7%), Book Chapters (16.9%), and Journal Articles (18.8%). Over 40% of the publications discovered are in Computer Science, with 25% in Engineering. As per Table 1, the most common source for cyber warfare research publications by African authors is the European Conference on Information Warfare and Security and International Conference on Information Warfare and Security. Since the names of sources change every year, especially on conference proceedings, some of the data in the Table 1 below may not be complete.

Table 1: Most relevant sources of cyber warfare research in Africa

Sources	Articles
EUROPEAN CONFERENCE ON INFORMATION WARFARE AND SECURITY	23
INTERNATIONAL CONFERENCE ON CYBER WARFARE AND SECURITY	22
LECTURE NOTES IN NETWORKS AND SYSTEMS	5
INFORMATION SECURITY FOR SOUTH AFRICA	3
EUROPEAN CONFERENCE ON INFORMATION MANAGEMENT AND EVALUATION	3
ACM INTERNATIONAL CONFERENCE PROCEEDING SERIES	3
COMPUTER AND INFORMATION SECURITY HANDBOOK	3
LECTURE NOTES IN COMPUTER SCIENCE	3

Figure 3 below shows that since 2005, the African continent has been slowly increasing research on cyber warfare, starting from below 5 articles per year in 2005 to 15 articles (on average) in 2023. The growth rate of publications on cyber warfare in Africa is estimated at 17.7% over the past 18 years.

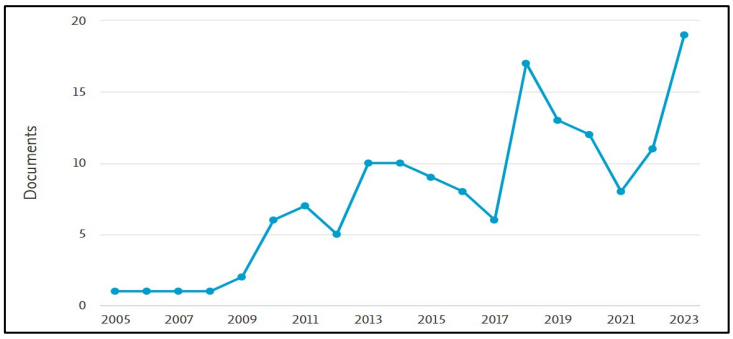


Figure 3: Cyber warfare research publication trends since 2005

4.2 Keywords, Trends, and Thematic Analysis

Relevant keywords provided by authors of the papers and occurred for more than 5 times in the Scopus core database were considered in the final analysis. Based on the 154 documents, 77 items met the threshold and a total of 8 clusters were generated with a total link strength of 404 based on the analysis done on VOSviewer. As may be noted in Figure 4 below, cyber warfare was a dominant cluster and links to a series of topics including cyber-crime, cyber defence, information operations, cyber terrorism, cyber counter-intelligence, cyber-attack, and others.

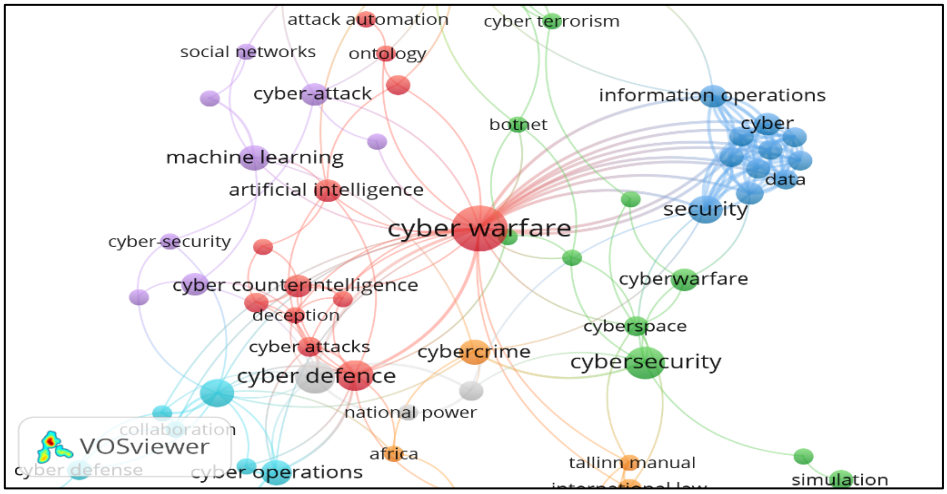


Figure 4: Keyword analysis on cyber warfare research in Africa

4.3 Authors, Collaboration, and Publication Impact Analysis

Figure 5 below shows the top ten authors in cyber warfare research in Africa over the past 18 years. Author "van Niekerk, B" has published, according to the data collected and filtering, 18 articles, followed by "Leenen L" with 14 articles. Both these authors have been productive in the cyber warfare domain since 2010.

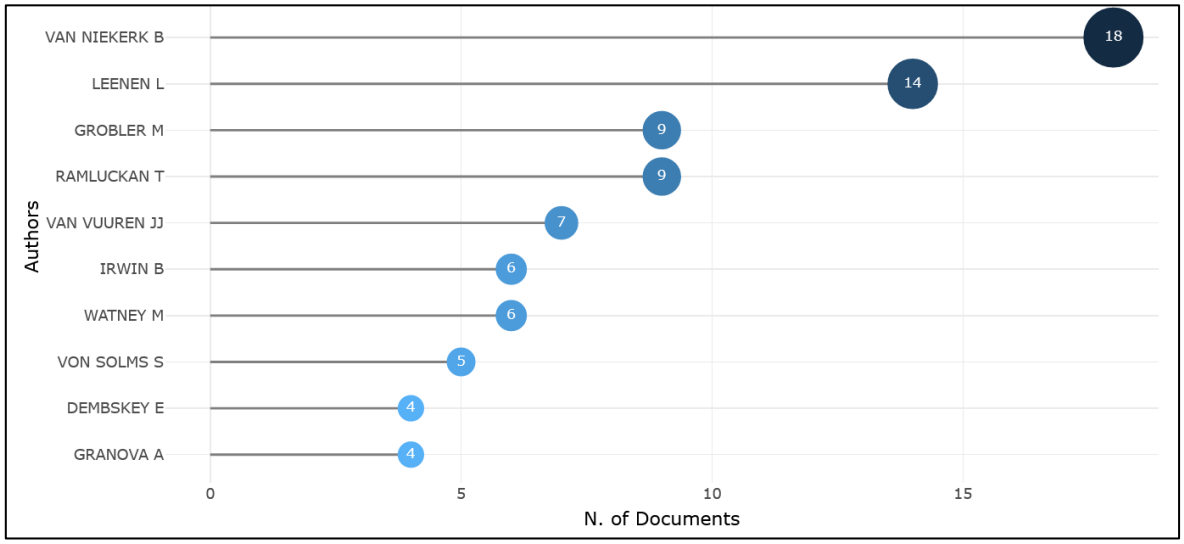


Figure 5: Author's productivity on cyber warfare research over time

In terms of collaboration networks amongst authors in Africa, the networks are quite diverse, albeit not strong. Author "Leenen, L" had a diverse collaboration network with different cohorts of researchers including those in the military environment.

The impact of cyber warfare research in Africa as depicted in Figure 6 and was at an average citation of 3.27 per document by end of year 2022. The top author had the highest h-index of five on cyber warfare research. These results are supported by the research published by Kyrdoda, et al., (2023) where African countries do not appear in the list of top countries with high-impact cybersecurity research.

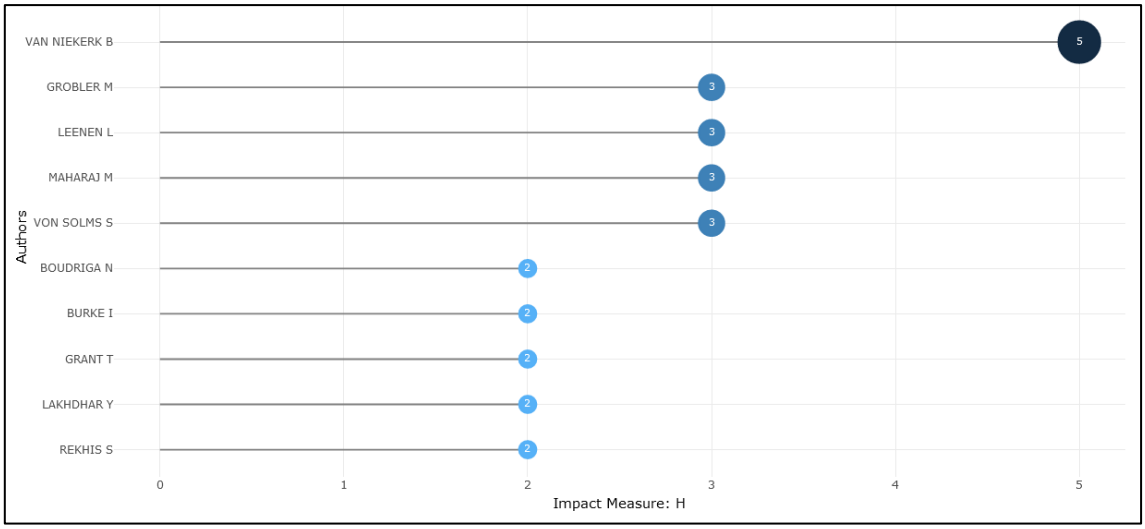


Figure 6: Publication impact factor analysis on cyber warfare research

Figure 7 shows the research collaborations through co-authorships between African countries.

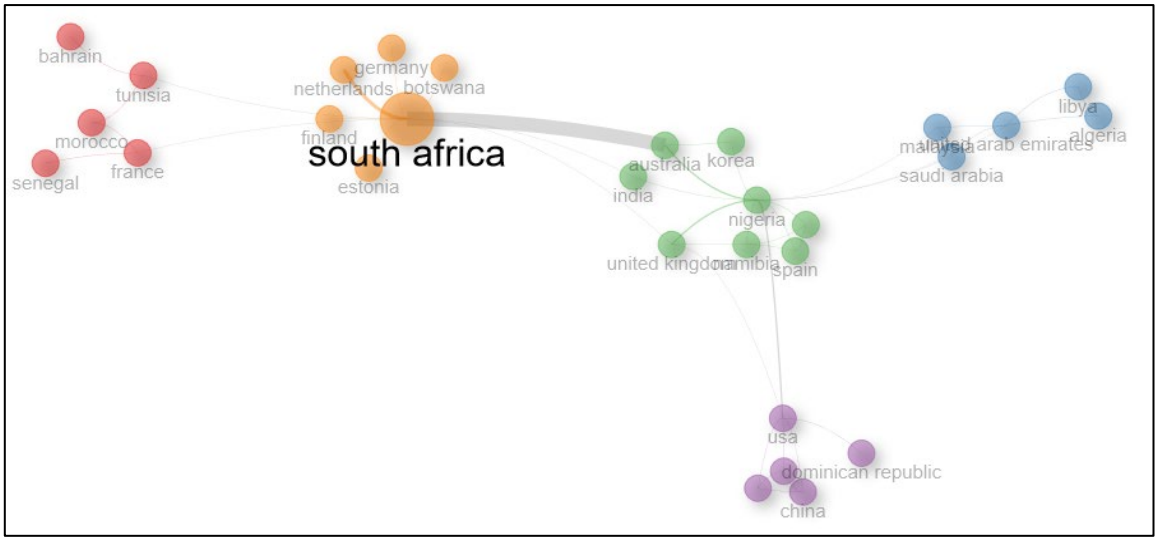


Figure 7: Research collaborations on cyber warfare research in Africa

The active countries (>5 articles) on cyber warfare research in Africa that were found in the collected data are: (1) South Africa, (2) Nigeria, (3) Morocco, (4) Tunisia, and (5) Algeria. The bibliometric analysis revealed that research collaborations within the cyber warfare domain in Africa tends to be with countries outside the region. And that there are limited collaborations between African authors on this subject matter as demonstrated in Figure 7.

4.4 Research Organisations and Funding Analysis

The organisations that are involved in the publications of cyber warfare research in Africa are mostly based in South Africa (see Figure 8) with the University of Carthage in Tunis and Covenant University in Nigeria, and Abdelmalek Essaadi University in Morrocco being the only institutions outside South Africa featuring in the top 10.

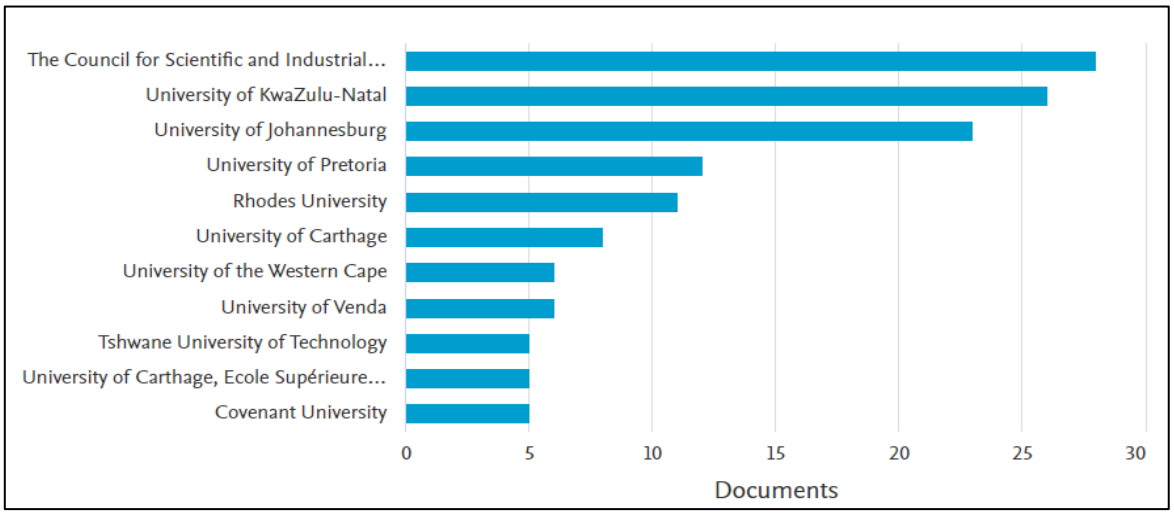


Figure 8: Research affiliations on cyber warfare in Africa

The organisations that are involved in the publications of cyber warfare research in Africa are mostly based in South Africa (see Figure 8) with the University of Carthage in Tunis and Covenant University in Nigeria, and Abdelmalek Essaadi University in Morrocco being the only institutions outside South Africa featuring in the top 10.

The analysis indicated that cyber warfare research funding in Africa is limited, and was generally supported by the National Research Foundation, South African Department of Defence, Armscor Ledger Program, Security and Software Engineering Research Centre and in some instances by overseas sponsors such as Hague Centre for Strategic Studies between 2005-2023.

5. Comparative Analysis

To contextualize the cyber warfare research trends and landscape in Africa, a comparative analysis was also conducted. This was done by choosing the top five countries in the National Cyber Power Index (NCPI) (Voo, et al., 2022) and use the same keywords to extract high-level publication data based on "article titles, abstract, and keywords" and other indicators as presented in Table 2. In Africa, South Africa* was used for comparison purposes because it had close to 70% of the publications in cyber warfare research over the past 18 years as of November 2023.

Based on the comparative analysis, the US is dominating across all publication indicators. The cyber warfare research in the US, China and UK is driven by a diverse set of research, private, and public institutions, including the military. The research in this domain tends to be funded by different funding sources or agencies, mostly local funding agencies in the United States and China. In South Africa, it emerged from the analysis that all cyber warfare research is driven by research and academic institutions. Funding sources of cyber warfare research in Africa is also limited to few funding agencies.

In addition, developed countries tend to succeed in collaborating with other research authors across the globe. Thus, the USA has a large group of research affiliations involved in cyber warfare research due to local and international collaborations. This, in our view, ultimately translates to a higher research impact.

Table 2: Comparative analysis of countries in cyber warfare research

Country	Search results within all fields	Search results in abstracts	Search Results in keywords	Search Results in article title	Search Results in Article title, abstract, and keywords*
United States of America	9,966	1,691	1,012	741	2,192
Republic of China	3,483	416	233	73	463
Federation of Russia	811	155	104	36	206
United Kingdom	3,352	353	198	187	489
Australia	1,695	157	86	55	191
South Africa*	697	87	71	42	105

6. Discussion, Implications, and Limitations

The bibliometric analysis study has yielded valuable insights into the landscape of cyber warfare research and development in Africa. Upon examining the sample of documents retrieved through the Scopus search query, it becomes evident that research activity in this domain is relatively limited. Only a total of 154 documents were sampled from African countries collectively, in contrast to 2,192+ documents for the United States over an 18-year period. This translates to approximately 9 research publications per year for Africa, compared to 122 articles from the United States and 27 articles from the UK per year.

Consolidated efforts are recommended between research affiliations and authors to elevate the rate of cyber warfare research. This could be achieved by establishing African publication forums and conferences specifically focusing on cyber warfare. According to the analysis, most of the African cyber warfare research is published in overseas-based research platforms.

Furthermore, the comparison reveals that cyber warfare research in Africa constitutes only 7% of the entire body of research in the United States and approximately 75% of the research output in the Russian Federation (refer to Table 2: "article title, abstract, and keyword" column for details). This significant discrepancy underscores the need for increased attention and investment in cyber warfare research and development within the African context.

The analysis also suggests that only two authors in the entire African continent have been consistently publishing in cyber warfare research over the past 18 years using the sampled data from Scopus, and both researchers are based in South Africa. This also translates to the limited publication impact of cyber warfare research in Africa,

with only 2.1 citations per published document, while in countries such as the USA, this can even reach 40+ citations per cyber warfare research publication.

An observation from the research findings is that African countries do not adequately collaborate or co-author research work in this domain. This limitation hinders the growth of cyber warfare research in Africa as well as its associated impact. It is recommended that African researchers need to break down barriers and intentionally collaborate on cyber warfare research with the public and private entities, ultimately assisting state entities in establishing and operationalizing cyber warfare capabilities.

Research findings indicate that countries that are highly productive in cyber warfare research are supported by a large cohort of active research affiliates, engage in multi-disciplinary collaborations, work with national security and defence agencies, and receive national research sponsorship. Therefore, it is recommended that for African states to grow their cyber warfare research activities, investments need to be made by both the public and private sectors.

On a positive note, the primary topics of research on cyber warfare in Africa are comparable to those in developed nations, suggesting that African nations are conducting relevant research in the third domain, albeit in a limited scope.

Limitations identified in this research study include the fact that Scopus indexing does not cover the entire research universe, potentially understating the African perspective. Keyword searches also have limitations that may impact the extent of research on cyber warfare in Africa. Furthermore, the results may be negatively affected by diverse naming conventions in article titles, authors' affiliations, publication platforms, and other factors. The bibliometric analytical tools are also not yet mature enough to automatically eliminate not so obvious duplicates due to varying naming conventions, requiring manual intervention, which may introduce errors and bias in the analysis.

7. Conclusion and Future Research

The research and development of cyber warfare capabilities in Africa are currently limited. However, South Africa stands out as a leader in this field, actively contributing to the global community and ranking twelfth in the number of cyber warfare research publications, with the USA leading all countries. This research highlights the potential of bibliometric analysis in revealing trends and patterns within a selected research domain. The findings emphasize that bibliometric data is rich and can offer valuable insights for defining research directions and supporting policy formulations. Nevertheless, it is essential to acknowledge the challenges associated with retrieving data for research from databases like Scopus. To address some of these limitations, a comparative analysis was employed. It is important to note that these limitations are not exclusive to research in Africa but are widespread. Therefore, this research contributes not only to the cyber warfare research landscape in Africa, but also to the broader global context. In future research, it may be worthwhile to focus on studies indexed Web of Science and Scopus to understand the trends broadly. It may also be useful to consider using Google Scholar for a bibliometric analysis as it tends to index majority of publications. Lastly, studying the mandates and profiles of research affiliations involved in cyber warfare research could be helpful for African countries to improve cyber warfare research.

Acknowledgement and Declaration

This work received support from the Council for Scientific and Industrial Research in South Africa. Generative AI in Google Bard and ChatGPT were utilized on limited basis for language editing, sentence construction in certain sections of this document. All sources employed for this research are duly cited below.

References

Amoah, M., 2019. Sleight is right: Cyber control as a new battleground for African elections. *African Affairs,* 119(474), pp. 68-89.

Aria, M. & Cuccurullo, C., 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics,* 11(4), pp. 959-975.

Binnendijk, A. L. et al., 2019. *Operationalizing cyberspace as a military domain.* 2019 ed. Santa Monica: Rand Corporation.

Blair-Walcott, K., 2023. Comparative Analysis. In: J. M. Okoko, S. Tunison & K. D. Walker, eds. *Varieties of Qualitative Research Methods.* s.l.:Springer, Cham, pp. 79-84.

Chen, C., 2014. *The citespace manual,* Drexel: College of Computing and Informatcs.

Colarik, H. D., 2017. The Hierarchy of Cyber War Definitions. In: G. Waag, M. Chau & H. Chen, eds. *Intelligence and Security Informatics.* s.l.:Springer, Cham.

Donthu, N. et al., 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research,* 133(2021), pp. 285-296.

Fielder, J. D., 2021. Cyber security in Kenya: Balancing economic security and internet freedom. In: *Routledge Companion to Global Cyber-Security Strategy.* s.l.:Routledge, pp. 543-552.

ITU, 2021. *Global Cybersecurity Index 2020.* [Online] Available at: https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2021-PDF-E.pdf [Accessed 24 November 2023].

Kyrdoda, Y., Marzi, G., Dabić, M. & Daim, T. U., 2023. Cybersecurity Technology: An Analysis of the Topic from 2011 to 2021. In: *Cybersecurity. Applied Innovation and Technology Management..* s.l.:Springer, Cham.

Lewis, D. M. & Alpi, K. M., 2017. Bibliometric Network Analysis and Visualization for Serials Librarians: An Introduction to Sci2. *Serials Review,* pp. 239-245.

Lukin, K., 2019. Russian Cyberwarfare Taxonomy and Cybersecurity Contradictions Between Russia and EU: An Analysis of Management, Strategies, Standards, and Legal Aspects. In: *National Security: Breakthroughs in Research and Practice.* s.l.:IGI Global.

Microsoft, 2021. *Use the Analysis ToolPak to perform complex data analysis.* [Online] Available at: https://support.microsoft.com/en-us/office/use-the-analysis-toolpak-to-perform-complex-data-analysis-6c67ccf0-f4a9-487c-8dec-bdb5a2cefab6 [Accessed 25 November 2023].

Mtsweni, J. & Thaba, M., 2021. Building an Integrated Cyber Defence Capability for African Missions. *Journal of Information Warfare,* 21(1), pp. 17-34.

Nobanee, H. et al., 2023. Bibliometric analysis of cybercrime and cybersecurity risks literature. *Journal of Financial Crime.*

Rahim, N., 2021. Bibliometric Analysis of Cyber Threat and Cyber Attack Literature:Exploring the Higher Education Context. In: M. Sarfraz, ed. *Cybersecurity Threats with New Perspectives.* Kuwait: IntechOpen.

Thaba, M. & Mtsweni, J., 2023. *Developing Robust Cyber Warfare Capabilities for the African Battlespace.* Athens, Greece, European Conference on Cyber Warfare and Security.

The Times of Israel, 2023. *Israel-linked group claims cyberattack that shut down 70% of Iran's gas stations.* [Online] Available at: https://www.timesofisrael.com/israel-linked-group-claims-cyberattack-that-shuts-down-70-of-irans-gas-stations/ [Accessed 21 December 2023].

Tse, J. et al., 2015. A bibliometric analysis of privacy and ethics in IEEE Security and Privacy. *Ethics and Information Technology,* 17(2015), pp. 153-163.

van Eck, N. J. & Waltman, L., 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics,* 538(84), p. 523.

Voo, J., Hemani, I. & Cassidy, D., 2022. *National Cyber Power Index 2022,* Cambridge: Belfer Center for Science and International Affairs, Harvard Kennedy School.

Willet, M., 2023. The Cyber Dimension of the Russia–Ukraine War. *Survival: October-November 2022,* Issue Apr 2023, pp. 7-26.

Systematic Review to Propose a Blockchain-Based Digital Forensic Ready Internet Voting System

Edmore Muyambo[1] and Stacey Omeleze Baror[2]
[1]Digital Forensic Research Group, Dept of Computer Science
[2]University of Pretoria, South Africa

Edmore.muyambo@tuks.co.za
Stacey.baror@cs.up.ac.za

Abstract: The ballot paper-based voting system has a high risk of data manipulation and vote tampering due to a lack of immutability, transparency and privacy. This systematic review is conducted with the intention of proposing a digital forensic-ready internet voting process to mitigate issues of vote rigging and vote fraud. The review focused on current and up-to-date literature. Publications that are out of this date range were ignored and considered stale or irrelevant. We extracted and reviewed publications with either or all the following keywords: "digital forensic", "internet voting", "e-voting" and "blockchain". A total of thirteen databases were consulted. These include ACM, IEEE, Web of Science, Science Direct, Academic Search Complete, Access Science, ProQuest, Oxford Academy, Ingenta, Cambridge Core and Clarivate. From the inclusion list of one hundred and five (105) studies that were looked at in detail, twenty of them (i.e.,19%) covered blockchain technology. Two percent (2%) of them were focused on digital forensics in internet voting. The results herein were synthesised and presented in a qualitative methodology. The review shows that a secure and reliable digital forensic system could effectively mitigate vote rigging and fraud. Therefore, we proposed a blockchain-based digital forensic-ready internet voting system. The proposed system is beneficial to the electorate, election observers, electoral candidates, electoral administration bodies as well as the national law enforcement agencies.

Keywords: Digital forensics, Blockchain, Internet voting, Online voting, Cyber security

1. Introduction

With the rapidly growing Internet of Things, smart technologies are emerging. The emergency of smart technologies brings about the need for smart infrastructure. This holistically integrates to form smart cities. In smart cities, smart devices are interconnected to render municipal essential services such as water billing, electricity billing, traffic management as well as municipal e-voting processes amongst many other things. Therefore, it becomes essential that we conduct a systematic review of digital transactions, privacy and security. This gives us a clear understanding of the currently existing knowledge in the research field thereof.

Research has shown that many countries are still experiencing electoral issues, pre-election, during and post-election season (Gomes, Moura,2017). Most of these issues are linked to alleged vote rigging and vote fraud. These cases are mostly recorded in countries that are still using the ballot paper-based voting system (Smith,2016). This is due to the fact that the ballot paper-based system has a high risk of data manipulation and vote tampering (Singh, Wable, Khrose, 2021). The process involves several manual procedures. This exposes electoral data to human errors and malicious vote tampering practices that can subsequently result in vote fraud and vote rigging. As a result, electoral disputes usually ensue due to lack of immutability, transparency, privacy and data security in the process. However, some sources still argue that the traditional process is still better than the electronic means. Given all the above details, we decided to conduct an extensive systematic literature review to ascertain how the integration of internet voting and digital forensics can be used to mitigate vote rigging and vote fraud issues. The use of digital security technologies such as the blockchain is still relatively new in the industry. Therefore, the available systematic reviews in this research field are still quite few. Hence our motivation to conduct a review.

With reference to some studies in the domain, Blockchain has been identified as one such technological process that can be used to preserve data security, integrity and privacy within a set of interconnected devices and systems (Hjalmarsson, Hreioarsson et al, 2018). The studies highlighted that Blockchain can achieve the above-mentioned security benefits in a transparent yet uncompromisable manner (Ayed, AB, 2017). An attack or alterations to the secured datasets or digital transactions can be easily detected and flagged by the blockchain technology. However, the transparency characteristic risks the exposure of personal data to the public.

1.1 Research Question and Objectives

The research question in this study is: can digital forensics mitigate issues of vote rigging and vote fraud in internet voting processes? The objective of the literature revision is to systematically analyse current studies on the use of digital forensics backed by blockchain technology in a voting process.

2. Method

This review was done on publications that were peer-reviewed and published by accredited academic publishers. We reviewed publications that centered their research on digital forensics in internet voting in integration with blockchain technology. At least three of the main search keywords must exist in the publication. These are “digital forensic”, “internet voting” and “blockchain”. Publications that had some alternative search keywords such as “online voting” instead of “internet voting” were also taken up into the inclusion list; provided at least three of the main keywords (highlighted above) existed. Publications that had all the required keywords but were published some ten years ago were not reviewed. We consulted studies that were not more than ten years old. We were mainly interested in studies that investigated ballot paper-based voting systems, electronic-based voting systems and the blockchain technology. Other key screening criteria for our inclusion list were language, credibility and authenticity. Publications were supposed to be in English, peer-reviewed and published by internationally accredited publishers. For verifiability and referencing purposes, we reviewed only publications that have been stored in internationally recognized academic libraries and databases. We made very little to no reference at all to grey literature.

Key players in this research were voters, election candidates and the election administration board. The main aim of having a secure vote administration system is to protect voters’ choices.

2.1 Information Sources

Our information base was gathered from a total of thirteen databases. We also made use of referral websites. These include Google Scholar and the University of Pretoria’s online library. We narrowed the search to articles and documents published between the first of January 2015 and the thirty-first of March 2023. A full list of consulted databases and dates thereof is shown in Table 1 below.

Table 1: Consulted Databases

DATABASE	DATE SEARCHED
ACM	1 March 2023
IEE	1 March 2023
WEB OF SCIENCE	4 March 2023
SCIENCE DIRECT	4 March 2023
ACADEMIC SEARCH COMPLETE	4 March 2023
ACCESS SCIENCE	7 March 2023
PROQUEST	7 March 2023
OXFORD ACADEMY	9 March 2023
INGENTA	9 March 2023
OPEN GLOBAL TRUSTED	9 March 2023
CLARIVATE	9 March 2023
CAMBRIDGE CORE	11 March 2023
BRILL	11 March 2023

2.2 Search Technique

Several keywords were used to search for relevant literature. Only publications that contained the required keywords were considered for review. The main keywords used in the search console are Digital forensics, internet voting, blockchain, online voting and e-voting. A combination of keywords strategy was also used to retrieve relevant publications. A list of search keyword combinations is listed below.

- Digital forensic
- Digital forensic + internet voting

- Digital forensic + internet voting + blockchain
- Digital forensic + internet voting + blockchain + online voting
- Digital forensic + internet voting + blockchain + online voting + e-voting
- Digital forensic + internet voting + blockchain + online voting + e-voting + published date between 2015/01/01 and 2023/03/31

On the first search, "Digital forensic" keyword was used. This presented us with thousands of search results in return. This was quite a large number to reference and read through. We then made various search combinations to try and narrow the search results as well as to get more relevant studies for review. Different search permutations were used as shown above and the number of hits was recorded. After all keyword combination search was done, we applied a date filter to get newer publications. Newer publications are essential as they contain recent research knowledge and findings. Older publications introduce a risk of bias in a review as they were done based on old or now-irrelevant technologies.

After applying the date filter, the number of publications became even smaller. We then went on to manually read through titles and abstracts in order to extract more relevant publications for review. Figure 1 below shows the search strategy used.

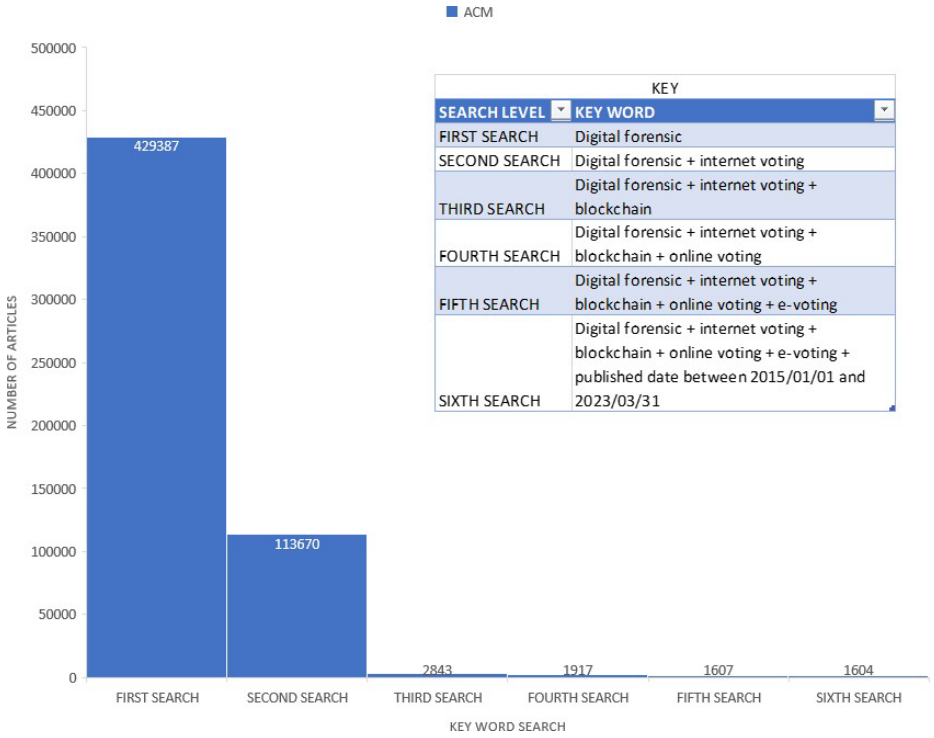


Figure 1: ACM Database Search Strategy

2.3 Selection Process

After applying the search strategy, publications that made it into the preliminary inclusion list were taken up for further screening. As a means of screening studies, we briefly read through titles, abstracts, introductions and the entire document. From this screening phase, we selected studies that would make it into the next phase of screening. The final screening was conducted by means of thoroughly studying the entire document. We believe that when the means of communication is not clear enough, it becomes difficult to understand the message being passed. Therefore, we also went on to exclude studies that we found linguistically unprofessional. The selection process was not done by groups or software but by a single reviewer.

2.4 Data Collection Process

Data collection was not done by means of some automated process or group collection but by a single reviewer. This was done through the means of manually reading published studies as well as studying reports, images, tables, graphs and numbers.

2.5 Data Items

The outcome domains in this review were:

- Inefficiency of the ballot paper-based voting system.
- Vulnerability of the ballot paper-based voting system to vote rigging.
- Efficiency of an online voting system based on blockchain technology.
- Importance of Blockchain technology in preserving data and digital transactions
- Importance of digital forensics in online voting processes

All the above outcomes are significant in the sought results. Results and conclusive findings for each outcome were accessed and considered. Proposed models and suggestive findings were also included in the inclusion list.

We collected information based on the year of publication, keyword match and literature relevance. The information collected was not country-specific, nor was it individual-specific. Any publication which met the screening criteria would make it into the inclusion list. Thus, our target publications were those that focused on voting, data encryption and digital forensics.

2.6 Study Risk of Bias Assessment

In this research, the risk of bias was conducted by assessing what is known thus far and comparing it with other studies' findings. If the comparison looked somewhat deviating from the known knowledge that has scientific proof, then a risk of bias would be flagged. We expect quantified findings to have a valid point of reference. Therefore, if there is a numerical representation in a study but there is no valid source of reference to it, we flagged this as a potential introduction of risk of bias in the review. This assessment was done by a single reviewer.

2.7 Effect Measures

From the one hundred and five (105) studied publications, ninety (90) of them i.e., 85.7% of them expressed clearly the vulnerability of ballot paper-based voting systems to vote rigging. They possess a high risk of undetectable vote tampering and untraceable vote manipulation activities. Of the ninety

(90) studies, eighty-five (85) of them i.e.,94.4% clearly stated how blockchain technology can be of great importance in the mitigation of vote rigging and vote fraud in online voting processes. Of these eighty-five publications that highlighted the importance of blockchain technology, only twenty (20) of them i.e., 23.5% referred to the importance of digital forensics in preserving data integrity in online transactions. There is still very limited knowledge, research and literature about digital forensics in online voting. The percentage of articles that spoke about digital forensics in online voting was 2%. Of the twenty (20) studies in our inclusion list that focused on blockchain technology, only 2% of these were focused on digital forensics in online voting.

2.8 Synthesis Methods

The coding used to determine the suitability of each study considered for synthesis in this research was:

- Online voting/Internet voting process
- Ballot paper-based voting process
- Blockchain technology
- Digital forensics
- Vote rigging

These codes of synthesis basically aimed to extract and understand information about the traditional voting system, its flaws and its strengths. The codes also aimed to extract information about online voting systems. It is essential that a study be conducted on the two identified voting systems. This would then give a clear indication of the strengths and weaknesses of each. Due to the high vote rigging allegations in a traditional voting system, this review targeted studies that highlight how the online-based voting system can be fortified and strengthened in order to mitigate such vote rigging. Blockchain technology was identified as one of the most reliable mechanisms that can be used to preserve online transactional data integrity, privacy and security. Hence the need to have blockchain studies in the inclusion list. The implementation of blockchain technology is just a technical process. This process gives a platform for the digital forensic process to take place. Hence all studies that met the digital forensic inclusion criteria were considered for synthesis.

We used a graphical representation (histogram) for our digital search findings. The graphical representations are per search level and keyword used in each library or database. An example is shown in Fig 1 above.

2.9 Reporting Bias Assessment

Online voting is still in the infancy stages. Likewise, the onboarding of the blockchain technology in the process thereof. Therefore, research studies in the implementation of blockchain technology to mitigate vote rigging and vote fraud are still quite limited due to lack of literature and research in the faculty. Search results from the keyword "blockchain" were compared with the search results from the "ballot paper voting process". There was quite a huge difference in the number of articles returned. The former had quite a few results whilst the latter had a huge number of studies returned. This shows that there is a gap of bias due to the lack of sufficient reference literature in the past. This assessment was evident and therefore did not require any collaboration or group effort. Hence, only a single reviewer contacted the assessment.

2.10 Certainty Assessment

From the search conducted, we derived a degree of certainty and confidence in the search results after assessing their consistency. The physical analysis was used to identify and separate isolated results from the popular ones. This assessment was carried out by only one reviewer.

3. Results

3.1 Study Selection

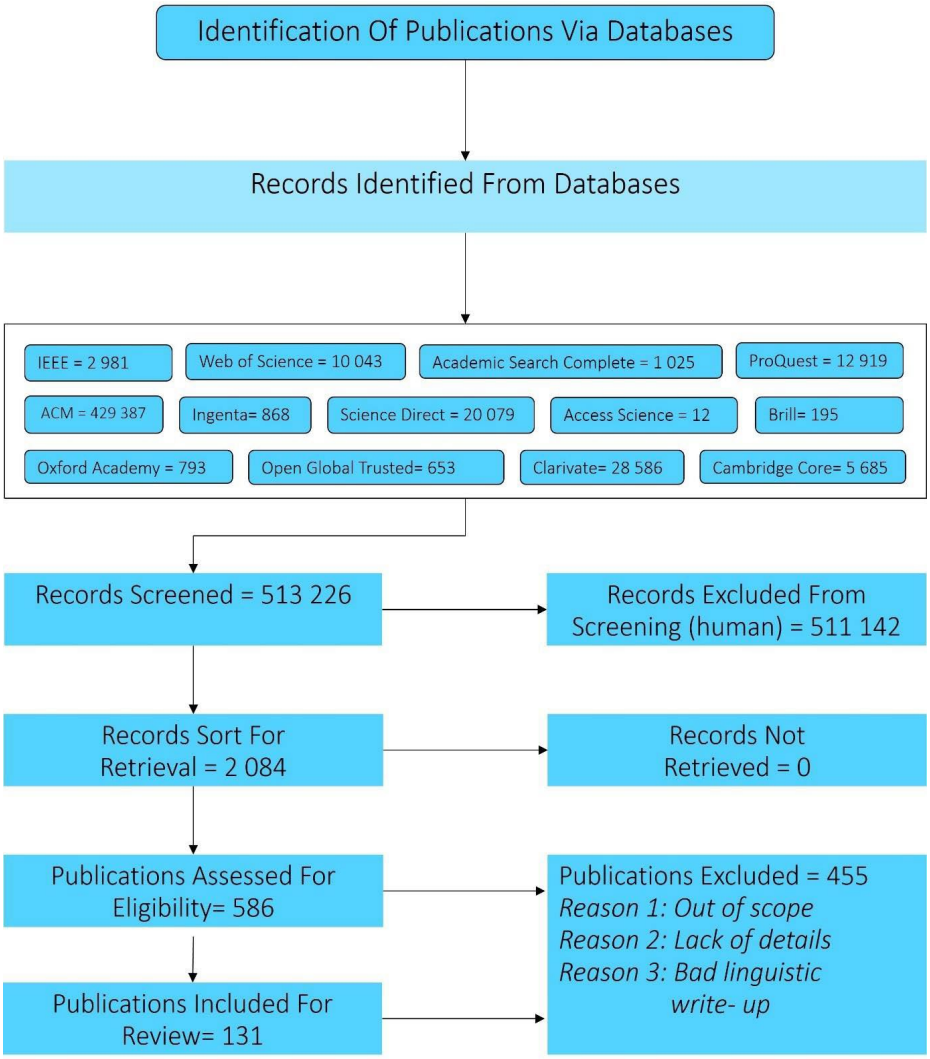


Figure 2: Search Results-Selection Criteria

As depicted in Fig 2 above, our initial search criteria gave us 513,226 records. These were then screened against our inclusion list criteria. The next screening phase excluded 511,142 records. This was done through the application of search filters such as keywords and year of publication. Of the 2 084 publications that survived the previous screening phase, 586 of them made it to the next round of screening. These were then manually assessed by physically reading through titles, abstracts and introductions. After this exercise, only 131 records made it to the next stage.

Of the 131 records from the previous screening stage, 17 were excluded. To prevent our paper from spanning too many pages, we only listed five of the seventeen excluded studies. These are shown in Table 2 below. In a nutshell, most of them were technically irrelevant to our review field, some lacked scientific detail and relevance, and some were just not properly written up grammatically. It was difficult to understand the concept and knowledge shared due to poor grammar or a lack of standard scientific writing.

Table 2: The Exclusion List

Author	Year	Tittle	Exclusion Reason
Neloy, Mohammad Nabiluzzaman; Wahab, Md. Abdul; Wasif, Sheikh; All Noman, Abdulla	2023	A remote and cost- optimized voting system using blockchain and smart contract	Does not cover digital forensic scope
Chaabane, Faten; Ktari, Jalel; Frikha, Tarek; Hamam, Habib	2022	Low Power Blockchained E-Vote Platform for University Environment	Limited scope
Vatsa, Avimanyou	2021	BCT-Voting : A Blockchain Technology Based Voting System BCT-Voting : A Blockchain Technology Based Voting System	Does not cover digital forensic scope
Dimitri, Nicola	2022	Quadratic Voting in Blockchain Governance	Too technical and lacks clear linguistic literature
Desai, Shreya; Desai, Shreya; Desai, Shreya; Li, Zhigang; He, Jing; Xu, Xiaohua	2019	Untampered electronic voting in entertainment industry: A blockchain-based implementation	Outside the required scope

3.2 Study Characteristics

Table 3: Included Study Citation

Author	Year	Tittle	Study Area
Li, Meng; Lal, Chhagan; Conti, Mauro; Conti, Mauro	2021	LEChain: A blockchain- based lawful evidence management scheme for digital forensics	Blockchain and Digital Forensic
Ryu, Jung Hyun; Sharma, Pradip Kumar; Jo, Jeong Hoon; Jo, Jeong Hoon	2019	A blockchain-based decentralized efficient investigation framework for IoT digital forensics	Blockchain and Digital Forensic
Koç, Ali Kaan; Yavuz, Emre; Çabuk, Umut Can; Çabuk, Umut Can	2018	Towards secure e-voting using ethereum blockchain	E-voting and Blockchain
Englbrecht, Ludwig	2020	A privacy-aware digital forensics investigation in enterprises	Digital Forensics
McCorry, Patrick; Mehrnezhad, Maryam; Toreini, Ehsan; Shahandashti, Siamak F; Hao, Feng	2021	On Secure E-Voting over Blockchain	E-voting and Blockchain

Author	Year	Tittle	Study Area
Gomes, Alexandre	2017	Blockchain Voting and its effects on Election Transparency and Voter Confidence	Blockchain
Huang, Jun; He, Debiao; Obaidat, Mohammad S.; Vijayakumar, Pandi; Luo, Min; Choo, Kim Kwang Raymond	2021	The Application of the Blockchain Technology in Voting Systems	Blockchain and Voting
Rabia, Fatih; Sara, Arezki; Sara, Arezki	2021	A survey on e-voting based on blockchain	Blockchain and E-Voting
Vladucu, Maria Victoria; Dong, Ziqian; Medina, Jorge; Rojas-Cessa, Roberto	2023	E-voting Meets Blockchain: A Survey	Blockchain and E- Voting
Shah, Akhil; Sodhia, Nishita; Saha, Shruti; Banerjee, Soumi; Chavan, Madhuri	2020	Blockchain Enabled Online-Voting System	Blockchain and Online Voting
Anitha, V; Marquez Caro, Orlando Juan; Sudharsan, R.; Yoganandan, S.; Vimal, M.	2023	Transparent voting system using blockchain	Blockchain
Mizzi, Jerome; Inguanez, Frankie	2018	Blockchain based E-Voting System	Blockchain and E-Voting
K., Edison; B., Venansius	2022	Blockchain Based Electronic Voting Protocol	Blockchain
Holmes, C	2022	How Blockchain Technology could support Democracy and E-Voting	Blockchain and E- Voting
Fatrah, Aicha; Kafhali, Said El; Haqiq, Abdelkrim; Salah, Khaled	2019	Proof of concept blockchain-based voting system	Blockchain
Hossain, Syeda Sumbul; Arani, Samen Anjum; Tanvir Rahman, Md; Bhuiyan, Touhid; Alam, Delwar; Zaman, Moniruz	2019	E-voting system using Blockchain technology	Blockchain and E- Voting
Abuidris, Yousif; Kumar, Rajesh; Wenyong, Wang	2019	A survey of blockchain based on e-voting systems	Blockchain and E- Voting
González, Camilo Denis; Mena, Daniel Frias; Muñoz, Alexi Massó; Rojas, Omar; Sosa-Gómez, Guillermo	2022	Electronic Voting System Using an Enterprise Blockchain	Blockchain and E- Voting
Lahane, Anita A.; Patel, Junaid; Pathan, Talif; Potdar, Prathmesh	2020	Blockchain technology based e-voting system	Blockchain and E- Voting
Leune, Kees; Punjwani, Jai	2021	Enhancing electronic voting with a dual-blockchain architecture	Blockchain and E- Voting

Author	Year	Tittle	Study Area
Bartolucci, Silvia; Bernat, Pauline; Joseph, Daniel	2018	SHARVOT: Secret SHARe-based VOTing on the blockchain	Blockchain
Alshehri, Ali; Baza, Mohamed; Srivastava, Gautam; Rajeh, Wahid; Alrowaily, Majed; Almusali, Majed	2023	Privacy-Preserving E- Voting System Supporting Score Voting Using Blockchain	Blockchain and E- Voting
Neloy, Mohammad Nabiluzzaman; Wahab, Md. Abdul; Wasif, Sheikh; All Noman, Abdulla; Rahaman, Mustafizur; Pranto, Tahmid Hasan; Haque, A. K. M. Bahalul; Rahman, Rashedur M.	2023	A remote and cost- optimized voting system using blockchain and smart contract	Blockchain and Smart Contract
Wahab, Yousif Mohammed; Ghazi, Alaan; Al-Dawoodi, Aras; Alisawi, Muthana; Abdullah, Sirwan Saber; Hammood, Layth; Nawaf, Asmaa Yaseen	2022	A Framework for Blockchain Based E- Voting System for Iraq	Blockchain and E- Voting

Table 3 above details a citation list of some of the included studies in this systematic review. The table details author(s), year of publication, publication title as well as the reviewed literature. In this research, the four main study areas of focus were Digital Forensics, Blockchain, Internet Voting and to some extent Cyber Security details.

3.3 Risk Of Bias in Studies

The risk of bias in this study was conducted by studying publications' context. In some cases, only the abstract was read in order to determine the risk of bias whilst in other cases both the abstract and introduction were read. However, in most cases the abstract, introduction and the entire publication's content were read through in order to get the logical flow of the study and its intentions.

3.4 Reporting Biases

Since electronic voting is still not widely adopted, literature relating to digital forensics in electronic voting is likewise very limited. From our research list, the main identified studies that specifically relate to blockchain-based digital forensics in internet voting are *LEChain: A blockchain-based lawful evidence management scheme for digital forensics*, *Using Blockchain Technology to Create a Secure Digital Voting System*, *Survey on Digital E-Voting System by using Blockchain Technology, Towards secure e-voting using Ethereum blockchain, On Secure E-Voting over Blockchain* and *Blockchain Voting and its effects on Election Transparency and Voter Confidence*. This is a very small number. Consequently, the knowledge acquired also spans a limited coverage. Hence a risk of results bias cannot be entirely ruled out.

3.5 Certainty of Evidence

To a greater extent, an internet-based voting system that is integrated into blockchain technology for digital forensics has shown to be more resistant to vote rigging than the ballot paper-based voting system. Of the total consulted studies, 95% of them concur that the ballot paper-based voting system is prone to human errors and electoral data manipulation which can subsequently lead to vote rigging and vote fraud.

From the research and information that we have gathered thus far, we can conclude that apparently there is no known system yet which can be used to effectively address issues of vote rigging and vote fraud in online voting. The currently available processes lack transparency, privacy, integrity, confidentiality and auditability. Therefore, there is a need to have a voting system that is non-comprised and protected. In case of any data tampering attempts, the activities must be swiftly and automatically detected for further investigations. As of now (at the writing of this paper), such a system does not exist.

3.6 Proposed System-Digital Forensic Readiness Model for Internet Voting

In their study, (Muyambo, Omeleze Baror, 2023) proposed a blockchain-based system that could be used to mitigate vote rigging in Internet voting systems. To mitigate issues of vote rigging and vote fraud, this study

introduces the Digital Forensic Readiness Model for Internet Voting (DFRMIV). The model consists of four main layers described below, namely, the vote casting layer, digital forensic data store, electoral board data store and the vote reconciliation layer. The main components of the DFRMIV's model are the voting process, electoral board vote processing, digital forensic vote processing and the vote reconciliation process as shown in Figure 4 below. When a vote is cast, it goes through the Blockchain encryption technology. This technology makes sure that the vote is secure, safe and auditable. After the encryption process, votes are simultaneously stored in the Electoral Board's data repository and the Digital Forensic data repository. The Digital Forensic Readiness process would then merge the two datasets and perform a vote reconciliation process. If there are vote irregularities or vote rigging attempts, the vote reconciliation process would flag this.

Figure 3 below shows a diagrammatic view of DFRMIV's technical details. The first block encloses the voting process, the second block encloses the data repository mechanism and the 3rd block encloses the vote reconciliation process.

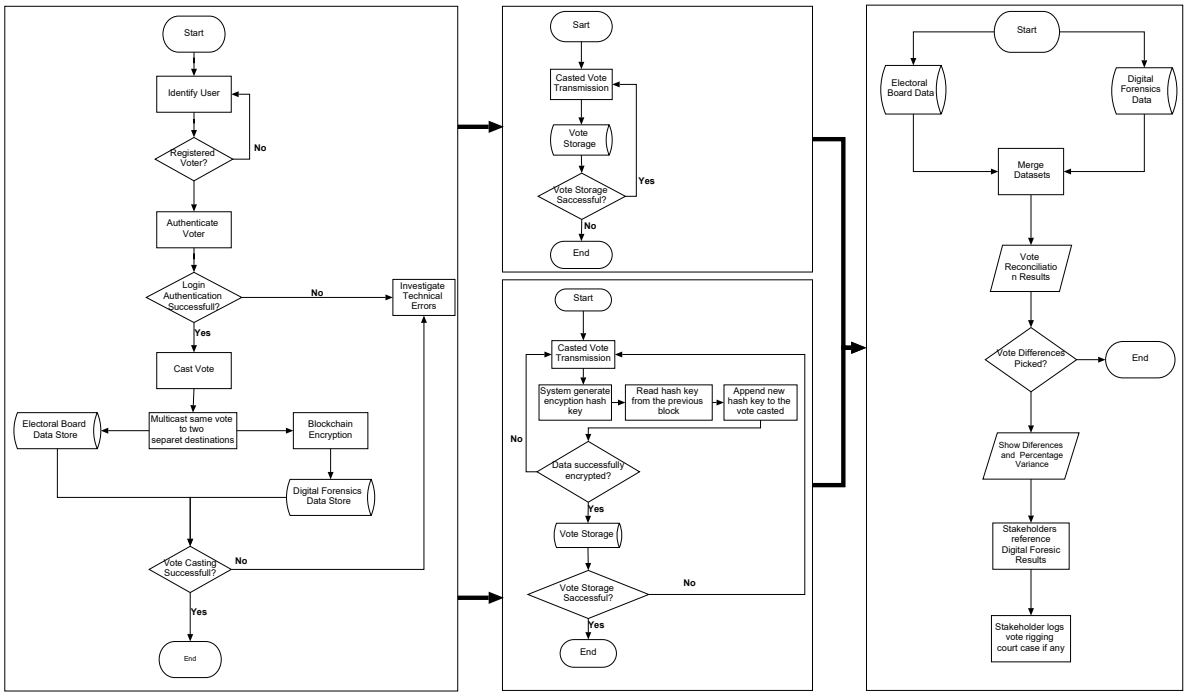


Figure 3: Proposed Blockchain-Based Digital Forensic Model Readiness Model for Internet Voting

3.6.1 Four main layers of the proposed system

- Vote Casting

This is the starting point of the election process. Before vote casting, every voter is required to register to vote. On the Election Day system users would log into the voting system for the vote casting process to begin. Once the user has been successfully verified against the voter's roll and authenticated, they can proceed to the vote-casting phase. The user chooses a candidate of their choice and submits the selection. When the submission process is completed successfully, the user will be automatically logged out. They will be presented with a confirmation of success and a reference number.

- Digital Forensics Data Store

When votes are cast, they are simultaneously distributed into two data repositories which are, the Digital Forensic Data Store and the Electoral Board Data Store. The Digital Forensic Data Store keeps raw data of votes cast. This dataset is used for digital forensic investigations i.e., vote rigging detection. This layer is not government-owned, it is owned by independent service providers.

- Electoral Board Data Store

Electronic votes cast are simultaneously captured into the digital forensic data store and the electoral board data store. This means by default the two data stores should contain exactly the same number of votes. If at any point there is a discrepancy in the number of votes between the digital forensics data store and the

electoral board data store, a possible vote-rigging activity could have occurred. The electoral board data store is used by the Electoral Board for vote counting, auditing and results presentation.

- Vote Reconciliation

This is a Digital Forensic Readiness component. Datasets from both data stores are merged and compared against each other. A difference between the datasets would mean a possible vote-rigging activity has occurred. The digital forensic readiness model for internet voting would automatically detect this and flag all the numeric deviations in detail.

3.6.2 Current state of work

In their study, Vatsa A and Raikar D highlighted the need to have a reliable blockchain-based online voting process to mitigate vote manipulation in an election process (Vatsa, 2021). The authors proposed a five-layer Blockchain Technology based Voting system (BCT-Voting) intending to mitigate vote fraud and data tampering. However, as we studied the author's proposal we found out that this system has several shortcomings. Amongst them is the ability to allow voters to alter their votes after casting. We believe this opens a room for vote tampering as well as vote manipulation. Voters must be able to cast their vote only once, without room to amend the vote thereafter. We also learned that BCT-Voting allows system administrators direct access to plain personal details in the database. Based on the principles of data privacy and cyber security, we hereby argue that this proposed system does not adequately meet the data privacy and voter secrecy checkpoints. Hence our motivation to propose a DFRMFIV system that cushions the shortcomings of the current proposals in the faculty.

3.6.3 DFRMIV evaluation

DFRMIV evaluation points

- Data accessibility – how safe and secure is the data access process? Is access safely granted and maintained?
- Data integrity – the integrity of the dataset in an end-to-end process. The dataset should be the same throughout the process and across systems. It should not be easily altered or manipulated without detection or trace.
- Data verifiability – data and digital transactions must be verifiable at any point in the process. Source and destination datasets should always be aligned. Misalignments should be accounted for (either in audit reports or any other data reconciliation processes).
- Data auditability – the ability to detect and reflect data changes in the process should be met. All transactions must be clearly recorded and the audit records must be easily accessible anytime by stakeholders.
- Digital forensics – datasets and all associated transactions must be securely handled through some robust technologies such as the blockchain in order to provide accurate digital forensics.

4. Discussion

Since the study of digital forensics in internet voting is still relatively new, we were faced with an issue of limited literature for review. Some of the studies lack explicit details such as quantification and source referencing. Some studies based their facts on general knowledge and empirical assumptions. However, according to (Adiputra, CK et al, 2018), it is quite evident that our research findings are well aligned with the little-known knowledge thus far about blockchain-based digital forensics in internet voting. (Kshetri, Vosa, 2018) highlighted that the current ballot box voting system is highly associated with vote rigging and data fraud. (German, Serdult, 2017) highlighted that it is arguable that internet voting increases voter turnout. This we could not attest to as our findings were quite limited to this fact-finding aspect.

This review was not conducted by a team or group but by a single reviewer. This exposes the review process to a risk of bias and inconsistency due to a lack of peer review and discussion. However, even though this review was conducted by a single reviewer, source references were clearly highlighted and the content referenced was understandably explained. Due to a lack of sufficient studies in the research field, not so many studies of relevance were found in the searched databases. In order to enhance clear communication, we only referenced English written studies. In avoidance of outdated information, we only referenced studies that were published not more than ten years ago from the writing of this publication.

From this review, implications for practice and policy are that ballot paper-based voting systems are quite vulnerable to vote rigging and vote fraud.

5. Other Information

This review was conducted as a Masters research student project. The paper was presented at the European Conference on Cyber Security (ECCWS) on the 24th of June 2023 in Greece, Athens. The full paper: https://papers.academic-conferences.org/index.php/eccws/article/view/1186/1216 . The paper has since been published to the world of academics for referencing. The research was self- funded (student). No external financial institutes or donor contributions. Therefore, this review was not registered under any organization. The author does not have any interests to declare.

References

Abuidris, Y., Kumar, R. and Wenyong, W., 2019, December. A survey of blockchain based on e-voting systems. In *Proceedings of the 2019 2nd International Conference on Blockchain Technology and Applications* (pp. 99-104).

Abuidris, Y., Kumar, R. and Wenyong, W., 2019, December. A survey of blockchain based on e-voting systems. In *Proceedings of the 2019 2nd International Conference on Blockchain Technology and Applications* (pp. 99-104).

Adiputra, C.K., Hjort, R. and Sato, H., 2018, October. A proposal of blockchain-based electronic voting system. In *2018 second world conference on smart trends in systems, security and sustainability (WorldS4)* (pp. 22-27). IEEE.

Alshehri, A., Baza, M., Srivastava, G., Rajeh, W., Alrowaily, M. and Almusali, M., 2023. Privacy- Preserving E-Voting System Supporting Score Voting Using Blockchain. *Applied Sciences*, *13*(2), p.1096.

Anitha, V., Caro, O.J.M., Sudharsan, R., Yoganandan, S. and Vimal, M., 2023. Transparent voting system using blockchain. *Measurement: Sensors*, *25*, p.100620.

Ayed, A.B., 2017. A conceptual secure blockchain-based electronic voting system. *International Journal of Network Security & Its Applications*, *9*(3), pp.01-09.

Bartolucci, S., Bernat, P. and Joseph, D., 2018, May. SHARVOT: secret SHARe-based VOTing on the blockchain. In *Proceedings of the 1st international workshop on emerging trends in software engineering for blockchain* (pp. 30-34).

Chaabane, F., Ktari, J., Frikha, T. and Hamam, H., 2022. Low power blockchained e-vote platform for university environment. *Future Internet*, *14*(9), p.269.

Denis González, C., Frias Mena, D., Massó Muñoz, A., Rojas, O. and Sosa-Gómez, G., 2022. Electronic voting system using an enterprise blockchain. *Applied Sciences*, *12*(2), p.531.

Desai, S., Han, M., Li, L., Li, Z., He, J. and Xu, X., 2019, September. Untampered Electronic Voting in Entertainment Industry: A Blockchain-based Implementation. In *Proceedings of the 20th Annual SIG Conference on Information Technology Education* (pp. 166-166).

Dimitri, N., 2022. Quadratic voting in blockchain governance. *Information*, *13*(6), p.305.

Djanali, S., Nugraha, D.P., Studiawan, H. and Pratomo, B.A., 2018. Vote identification and integrity of ballot in paper-based e-voting system. *Electronic Government, an International Journal*, *14*(3), pp.240- 254.

Englbrecht, L. and Pernul, G., 2020, August. A privacy-aware digital forensics investigation in enterprises. In *Proceedings of the 15th International Conference on Availability, Reliability and Security* (pp. 1-10).

Fatrah, A., El Kafhali, S., Haqiq, A. and Salah, K., 2019, October. Proof of concept blockchain-based voting system. In *Proceedings of the 4th International Conference on Big Data and Internet of Things* (pp. 1-5).

Germann, M. and Serdült, U., 2017. Internet voting and turnout: Evidence from Switzerland. *Electoral studies*, *47*, pp.1-12.

Hjálmarsson, F.Þ., Hreiðarsson, G.K., Hamdaqa, M. and Hjálmtýsson, G., 2018, July. Blockchain-based e-voting system. In *2018 IEEE 11th international conference on cloud computing (CLOUD)* (pp. 983- 986). IEEE.

Hjálmarsson, F.Þ., Hreiðarsson, G.K., Hamdaqa, M. and Hjálmtýsson, G., 2018, July. Blockchain-based e-voting system. In *2018 IEEE 11th international conference on cloud computing (CLOUD)* (pp. 983- 986). IEEE.

Holmes, C., 2022. How Blockchain Technology could support Democracy and E-Voting. *The Journal of The British Blockchain Association*.

Hossain, S.S., Arani, S.A., Rahman, M.T., Bhuiyan, T., Alam, D. and Zaman, M., 2019, December. E- voting system using blockchain technology. In *Proceedings of the 2019 2nd International Conference on Blockchain Technology and Applications* (pp. 113-117).

Huang, J., He, D., Obaidat, M.S., Vijayakumar, P., Luo, M. and Choo, K.K.R., 2021. The application of the blockchain technology in voting systems: A review. *ACM Computing Surveys (CSUR)*, *54*(3), pp.1- 28.

Kshetri, N. and Voas, J., 2018. Blockchain-enabled e-voting. *Ieee Software*, *35*(4), pp.95-99.

Lahane, A.A., Patel, J., Pathan, T. and Potdar, P., 2020. Blockchain technology based e-voting system. In *ITM Web of Conferences* (Vol. 32, p. 03001). EDP Sciences.

Leune, K. and Punjwani, J., 2021. Enhancing Electronic Voting With A Dual-Blockchain Architecture. *Ledger*, *6*.

Li, M., Lal, C., Conti, M. and Hu, D., 2021. LEChain: A blockchain-based lawful evidence management scheme for digital forensics. *Future Generation Computer Systems*, *115*, pp.406-420.

McCorry, P., Mehrnezhad, M., Toreini, E., Shahandashti, S.F. and Hao, F., 2021. On secure e-voting over blockchain. *Digital Threats: Research and Practice (DTRAP)*, *2*(4), pp.1-13.

Moura, T. and Gomes, A., 2017, June. Blockchain voting and its effects on election transparency and voter confidence. In *Proceedings of the 18th annual international conference on digital government research* (pp. 574-575).

Moura, T. and Gomes, A., 2017, June. Blockchain voting and its effects on election transparency and voter confidence. In *Proceedings of the 18th annual international conference on digital government research* (pp. 574-575).

Muyambo, E. and Baror, S.O., 2023, June. Digital Forensic Readiness Model for Internet Voting. In *European Conference on Cyber Warfare and Security* (Vol. 22, No. 1, pp. 657-667).

Neloy, M.N., Wahab, M.A., Wasif, S., All Noman, A., Rahaman, M., Pranto, T.H., Haque, A.B. and Rahman, R.M., 2023. A remote and cost-optimized voting system using blockchain and smart contract. *IET Blockchain*, *3*(1), pp.1-17.

Neloy, M.N., Wahab, M.A., Wasif, S., All Noman, A., Rahaman, M., Pranto, T.H., Haque, A.B. and Rahman, R.M., 2023. A remote and cost-optimized voting system using blockchain and smart contract. *IET Blockchain*, *3*(1), pp.1-17.

Raikar, D. and Vatsa, A., 2021. BCT-Voting: A Blockchain Technology Based Voting System. In *The 27 th International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'21), July* (pp. 26-29).

Raikar, D. and Vatsa, A., 2021. BCT-Voting: A Blockchain Technology Based Voting System. In *The 27 th International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA'21), July* (pp. 26-29).

Ryu, J.H., Sharma, P.K., Jo, J.H. and Park, J.H., 2019. A blockchain-based decentralized efficient investigation framework for IoT digital forensics. *The Journal of Supercomputing*, *75*, pp.4372-4387.

Shah, A., Sodhia, N., Saha, S., Banerjee, S. and Chavan, M., 2020. Blockchain enabled online-voting system. In *ITM Web of Conferences* (Vol. 32, p. 03018). EDP Sciences.

Singh, S., Wable, S. and Kharose, P., 2021. A review of e-Voting system based on blockchain technology. *International Journal of New Practices in Management and Engineering*, *10*(04), pp.09-13.

Smith, R., 2016. Confidence in paper-based and electronic voting channels: evidence from Australia. *Australian Journal of Political Science*, *51*(1), pp.68-85.

Wahab, Y., Ghazi, A., Al-Dawoodi, A., Alisawi, M., Abdullah, S., Hammood, L. and Nawaf, A., 2022. A Framework for Blockchain Based E-Voting System for Iraq. *International Journal of Interactive Mobile Technologies*, *16*(10).

Wei, C.C.Z. and Wen, C.C., 2018. Blockchain-based electronic voting protocol. *JOIV: International Journal on Informatics Visualization*, *2*(4-2), pp.336-341.

Yavuz, E., Koç, A.K., Çabuk, U.C. and Dalkılıç, G., 2018, March. Towards secure e-voting using ethereum blockchain. In *2018 6th International Symposium on Digital Forensic and Security (ISDFS)* (pp. 1-7). IEEE.

Theory-Guided Feature Selection in Cybercrime Data Science

Rennie Naidoo and Shiven Naidoo
University of the Witwatersrand, Johannesburg, South Africa

Rennie.Naidoo@wits.ac.za
shiven.naidoo@students.wits.ac.za

Abstract: Cybercrime data science is being significantly hampered by the presence of 'noisy' features within vast and complex datasets. We draw from the theoretical insights of the behavioural sciences to propose a feature selection model to enrich and improve the value and interpretability of cybercrime intelligence datasets. We piloted our theory-guided feature selection approach on a subset of intelligence datafeeds provided by a global fraud and cybercrime tracking firm. The results of the proposed social influence feature selection model show significant improvement in the interpretability of the machine learning-based exploratory analysis and advanced visualization techniques in an experimental setting. The feature selection model yielded rich insights about cybercriminal psychological tactics from social engineering scam data and has potential applicability in the areas of cyberthreat response and cybersecurity awareness training. Our study shows the value of an interdisciplinary theory-guided approach to cybercrime data analytics that integrates scientific knowledge from the behavioural sciences and data science expertise. Our paper concludes by suggesting avenues for future research on theory-guided feature selection seeking to incorporate behavioural science knowledge in cybercrime data science. We intend to refine, automate, evaluate, and scale our model in future research to assess its effectiveness in producing insights about cybercriminal activities and informing decision-making in a naturalistic and real-time setting. In future research efforts, we aim to automate the encoding of features and apply a wider range of machine learning tools and evaluation metrics to extract more meaningful insights into cybercriminal psychological tactics. We also intend to refine our model on larger datasets to enhance its efficiency and responsiveness to real-time cybercrime data. We call on data scientists and cybercrime domain experts to work together to apply theory-guided feature selection to improve processes of knowledge discovery that enhance our cybersecurity capabilities.

Keywords: Behavioural science, Feature selection, Cybercrime data analytics, Cybersecurity, Theory-Guided data science

1. Introduction

We are drowning in information and starving for knowledge.

–Rutherford D. Roger

Recent advances in data science and big data analytics (BDA) holds a lot of promise in finding innovative ways to tackle the scourge of cybercrime (Kamenov, 2018; Rawat et al., 2021; Sarker et al., 2020). While cybersecurity units and their data science teams receive vast amounts of data from various external and internal sources to monitor and respond to cyberincidents and cyberthreats (Suraj et al., 2018; Wang and Jones, 2021), the noise content in this information can adversely affect the quality of analysis and lead to misleading conclusions if not properly addressed (Gupta and Gupta, 2019; Zhu et al., 2003). We agree with scholars that the noise challenges in these datasets are further exacerbated by the spurious relationships detected in the data-driven models that have little or no scientific basis (Zhu and Wu, 2004).

Despite rapid advancements in computing power, databases, statistical methods and algorithms that can process and analyse large volumes of structured and unstructured data (Chen et al., 2012; Hastie et al., 2001; McAfee et al., 2012), we agree with scholars that much of the popular hype about the 'end of theory' and 'the death of the scientific method' in the popular press as a consequence of data science has been hasty (Karpatne et al., 2017). The concept of theory-guided data science (TGDS) suggests a complementary approach that seeks to integrate scientific knowledge and data science thereby uniting the world of data scientists and domain experts (Bechor and Jung, 2019; Downton et al., 2020; Viaene, 2013). This approach values the existing concepts, models, measures, and hypotheses that form part of the existing scientific knowledge base that explains a phenomenon and examines their fit with the available data instead of over relying on often spurious correlations (Wagner and Rondinelli, 2016). At the same time, theory-based models are susceptible to making assumptions that oversimplify complex phenomena (Downton et al., 2020; Faghmous and Kumar, 2014). Notwithstanding, according to Karpatne et al. (2017) theory-guided data science (TGDS) "uses the unique capability of data science models to automatically learn patterns and models from large data, without ignoring the treasure of accumulated scientific knowledge." This necessitates more effective inter-disciplinary collaborations between data scientists and domain experts (Mao et al., 2019). Similarly, we also believe that collaborations that combine theory-based models and data science models can significantly expand the potential of generating useful insights from complex cybercrime intelligence datasets.

There is a tremendous opportunity to jointly apply theory-based models and data science models to analyse cybercrime intelligence datasets, to more effectively detect, prevent, and respond to cyber threats and cybercrime (Foroughi and Luksch, 2018). Given the deluge of cybercrime data and the high noise potential in these large, diverse and complex cybercrime intelligence datasets, traditional data science and big data analytics on their own are arguably not sufficient to provide reliable and valuable insights into decision making (Rassam et al., 2017). Scam data involving social engineering techniques, where scammers manipulate individuals into divulging sensitive information, contains a significant amount of noisy content that does not contribute to successful scams but forms a substantial part of the overall dataset (Gupta and Gupta, 2019; Naidoo, 2020). As a subcomponent of TGDS, theory-guided feature selection (TGFS) is the process of incorporating theoretical advances from the problem domain to select relevant features for TGDS models (Bhowmik and Bhowmik, 2022; Miao and Niu, 2016; Slawski et al., 2010). By leveraging existing theories, concepts, principles, and expert insights, behavioural scientists and data scientists can work together to apply theory-guided feature selection to reduce noise in complex cybersecurity datasets (Mao et al., 2019). For example, behavioural scientists can apply theories and principles to recognize the relevant features in scam content. However, discerning relevant features from vast and diverse cybercrime datasets to develop more effective countermeasures remains a major challenge (Heartfield et al., 2016). This pilot study embarks on the exploration of theory-guided feature selection techniques on a small subset of a cybercrime intelligence datafeed provided by a global fraud and cybercrime company tracking cybercrime incidents. More specifically, we assess the utility of social influence theory in filtering out noise and focusing on the relevant content of social engineering scams, streamlining the dataset for analysis and interpretation (Cialdini and Goldstein, 2004; Wright et al., 2014). Thus, this study addresses the following question: *How can the application of theory-guided feature selection techniques using social influence concepts and principles enhance the interpretability of cybercrime intelligence feeds containing social engineering content?*

This paper is organized as follows: Section 2 provides a brief literature review of the proposed social influence theory-guided feature selection approach, highlighting the key social influence concepts and principles used in our feature selection process. Section 3 details our pilot case study methodology for feature engineering the cybersecurity intelligence datafeeds, detailing our collaboration architecture and process for theory-guided feature selection. Section 4 presents the results of our exploratory data analysis (EDA) and visualization of the social influence features and discusses their implications. While the findings from these results are limited to a pilot study, they will serve as a foundation for more extensive research in the future. Section 5 concludes the pilot study, highlighting the contributions of theory-guided feature selection to the field of cybercrime analytics and avenues for future research. This study contends that theory-guided feature selection techniques, grounded in domain-specific knowledge of the behavioural sciences, are instrumental in advancing cybercrime data science.

2. Social Influence Theory-Guided Feature Selection

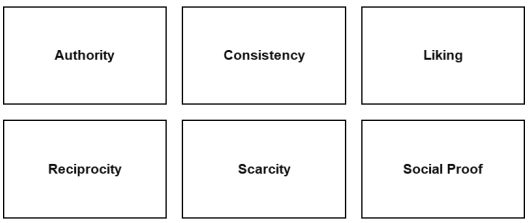


Figure 1: The Social Influence Feature Set

Cialdini's six persuasion principles, rooted in the psychology of compliance, serve as the primary knowledge base for the proposed theory-guided feature selection model, depicted as the social influence feature set in Figure 1. Drawing upon the concepts and principles of authority, consistency, liking, scarcity, reciprocity, and social proof (Cialdini and Goldstein, 2004; Wright et al., 2014), we craft our theory-guided data science model.

In Table 1 we show how our theory-guided feature selection process recasts the abovementioned social influence principles and definition as features and feature marker patterns respectively. The first feature is 'Authority,' and the feature pattern is where individuals tend to unquestioningly accept statements from perceived authorities (Milgram, 1963). 'Consistency' signifies the strong human inclination to align actions with commitments (Ferreira et al., 2015). 'Scarcity' heightens the value of items limited in availability, generating

increased demand. 'Reciprocation' captures the human tendency to feel obliged to return favours (Ferreira et al., 2015). 'Liking' highlights friendship pressures and people's readiness to assist friends, especially in online interactions, showcasing cooperative behaviour patterns (Ferreira et al., 2015). The last feature in our model, 'Social Proof' illustrates the impact of similar others on individual behaviour, leading to the adoption of similar actions or choices (Stajano and Wilson, 2011). Our goal is to enhance the interpretability of complex social engineering scams using the proposed social influence feature set model, to provide richer insights into cybercriminal tactics.

Table 1: Social Influence Feature Set Model with Feature Flags

Features	Feature Marker Patterns	Flag
Authority (A)	I follow the advice and obey the instructions of authority figures.	Present: (yes=1, no=0)
Consistency (C)	Once I have made a commitment, I do it.	Present: (yes=1, no=0)
Liking (L)	I like the agent therefore I am more inclined to believe them.	Present: (yes=1, no=0)
Reciprocity (R)	When the agent does me a favour, I feel inclined to return this favour.	Present: (yes=1, no=0)
Scarcity (S)	If the opportunity is for a limited time, I would act quickly to receive it.	Present: (yes=1, no=0)
Social Proof (SP)	If other people complied, I am more likely to comply.	Present: (yes=1, no=0)

3. Pilot Study: Social Influence Feature Engineering for Cybercrime Datasets

We used a small subset of data received from FraudWatch International's Cyber Intelligence Datafeed. These datafeeds were collected over a one-month period from March to April, 2020. Table 1 shows how the feature selection process began with the development of a coding template. The template served as a structured guide for identifying relevant features. This systematic approach not only ensured consistency in the analysis but also facilitated a deductive reasoning process, allowing for the targeted selection of features. Figure 2 shows how the researchers analysed photos and images iteratively. 185 unique documents were analysed. In the process of applying our theory-guided feature selection model, we implemented feature flags or binary indicators to operationalize the manually selected features. This critical step involved assigning a binary value of 1 (indicating 'yes') to each of the manually selected features, effectively flagging their presence within each scam. Conversely, a value of 0 (indicating 'no') was assigned to features not selected, creating a distinct flag variable for each underlying social influence principle (Li et al., 2018; Raschka et al., 2020). The authors independently coded the content and achieved consensus in case of any discrepancies. A cybersecurity expert, not familiar with project, acted as an independent auditor and reviewed the key categories and operational definitions, and provided reasonable verification of the accuracy of the coding procedure. The data underwent classification based on the specific online technologies affected, including but not limited to email, various social media platforms, video telephony and online chat services, cloud file hosting platforms, and streaming media services. Furthermore, categorization was performed according to the types of organizations subjected to impersonation attempts, encompassing entities such as banks, technology brands, product labels, and government agencies.



Code	Features	Binary Encoding
A	Authority	1
C	Consistency	0
L	Liking	1
R	Reciprocity	0
S	Scarcity	1
SP	Social Proof	1

Figure 2: Manually Selected Social Influence Features

Our collaboration architecture in Figure 3 provides an overview of the theory-guided feature selection components for analysing cybercrime datasets. The architecture typically begins with intelligence from both internal and external cybercrime feeds. This data is then subjected to data ingestion and preprocessing. The insights from both domain experts and data scientists are used for feature identification and engineering. Subsequently, the selected features are collated into a dataset, which then undergoes modelling and analysis using machine learning models and evaluation tools. The final step of the workflow involves generating documentation and reports. Throughout the process, there is a continuous cycle of monitoring and refinement, ensuring that the models remain updated and relevant. As illustrated, a collaborative platform and shared workspace serve as the backbone for theory-guided data science. Figure 4 depicts a phased approach to data analysis comprising three distinct phases. In Phase 1, the focus lies on preparing the feature-selected dataset. Phase 2 is the exploratory data analysis stage. In Phase 3, known as training and evaluation, which lies beyond the scope of this project, will involve the application of automated selection techniques to train the models.

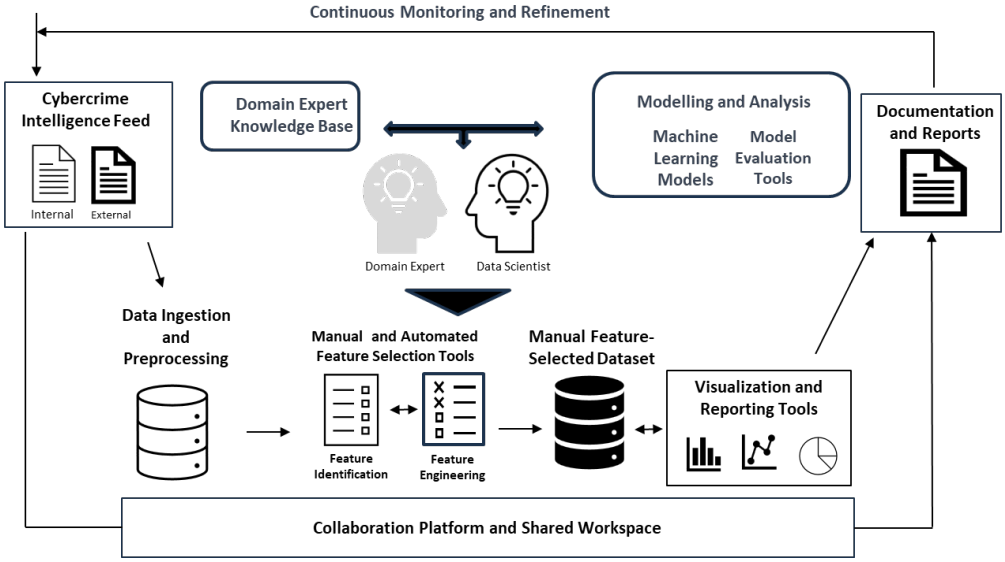


Figure 3: Collaboration Architecture for Theory-Guided Feature Selection

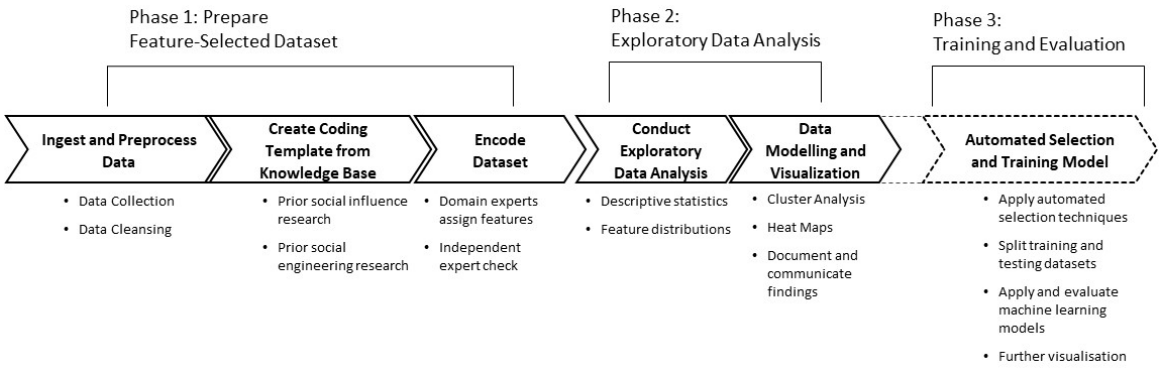


Figure 4: Manual Feature Selection Process

Python 3.0 was used for data processing, exploratory data analysis, clustering and visualizations. First, feature and scam category distributions were visualized using tools from the matplotlib library. The sklearn library was then used to perform and evaluate agglomerative clustering of the six social influence features. Next, the seaborn library was used to create a heatmap visualizing the usage of social influence features in different scam types. Lastly, the sklearn library was used to visualize the clustering of scam types based on their usage of social influence features using a t-SNE (t-distributed Stochastic Neighbor Embedding) map (Hastie et al., 2001; Raschka et al., 2020). We acknowledge that the lack of completeness of the dataset is a limitation of this study. In some cases, only phishing website information was provided and not the accompanying phishing email. Furthermore, it was difficult to assess to what extent the phishing email and phishing website were using the consistency and reciprocity tactics. Moreover, the archive documents were limited to active scams.

Notwithstanding, our approach ensured clear demarcation and identification of the social influence features, laying the groundwork for the analytical processes and results, discussed next.

4. Pilot Results and Discussion

4.1 Exploratory Data Analysis

Theory-guided EDA is important for understanding the distribution of characteristics for selected categories and features within a dataset. Figure 5 explores the distribution of 185 scams annotated according to two categories: Scam Types and Organization Types. Phishing was the dominant scam type, making up 60.9% of scams from the cybersecurity intelligence datafeed. The most common types of organizations impersonated by scammers included Social Networking Sites (SNSs) (39.1%), banks (23.9%), and technology organizations (21.2%). In contrast, there were relatively fewer scams impersonating government (8.2%) and intergovernmental (2.2%) organisations. The distributions of scam types and organisation types reveals important insights into scammer behaviour. Firstly, scammers tend to impersonate highly reputable institutions, as shown by the majority of scams impersonating banks, tech firms, government and intergovernmental organisations. Secondly, scammers may be intentionally impersonating larger international organisations in order to direct their scams to a wider, more global victimization base. This is supported by the large number of scams impersonating SNSs, technology organizations and banks.

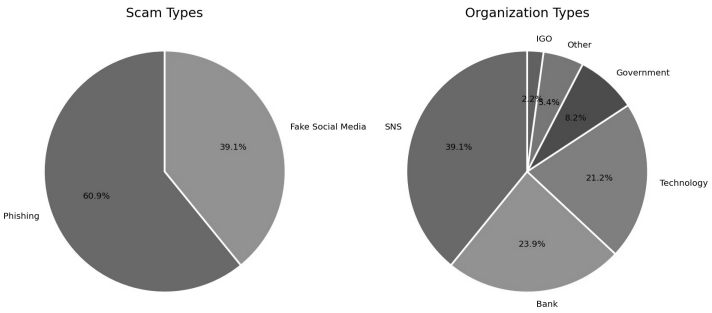


Figure 5: EDA; Distributions for Scam Categories

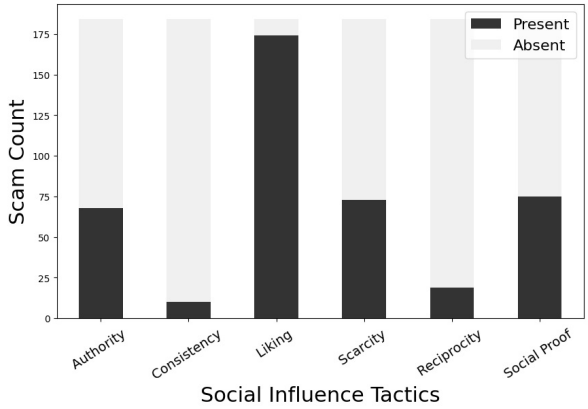


Figure 6: Prevalence of Social Influence Features in Scams

Figure 6 shows the prevalence of Cialdini's six social influence features in scams. Liking was the most employed tactic by scammers. Its widespread use suggests that scammers may primarily deceive users by impersonating people and institutions who are familiar or trustworthy, which is reflected in the types of organizations predominantly used by scammers in Figure 5. In contrast to liking, features like consistency and reciprocity were very rarely observed in scams. However, the cybercrime intelligence datafeed contained mostly active scams. Thus, there was little data showing scammer-user interactions, where these tactics are more likely to be employed. The remaining features, authority, scarcity and social proof, were all used in similar proportions but were still employed significantly less than liking.

4.2 Visualization of Social Influence Features

Visualizations were used to explore the relationships between selected categories and features (Embarak, 2018). Figure 7 shows the presence (1) or absence (0) of social influence features in scams. While liking seems

to be used ubiquitously, other features like authority, scarcity and social proof appear to be associated with specific scam types. For example, both authority and scarcity are commonly used in phishing scams. In contrast, social proof is commonly used in fake social media scams. Furthermore, consistency and reciprocity appear very rarely, but when they did appear it was almost always in phishing scams. When comparing the joint use of social influence features, it can be seen that authority and scarcity are often used together. Notably, social proof seems to be mutually exclusive with most social influence features, with the exception of liking.

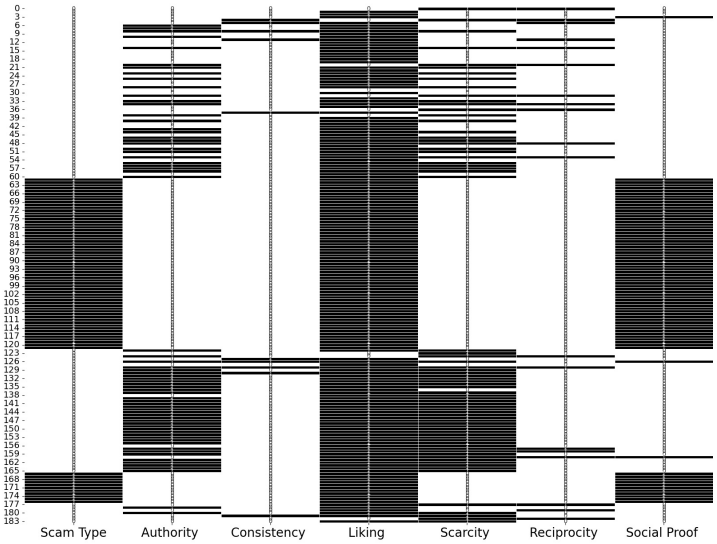


Figure 7: Heatmap showing Social Influence Features by Scam Types

Figure 8 below investigates the relationships between social influence features further using agglomerative clustering. The plot to the left shows the silhouette score for different n_cluster values when tuning the agglomerative clustering algorithm. Six clusters were used to optimise the clustering algorithm as it had the greatest silhouette score, suggesting the best cluster quality. To accommodate the data's binary encoding, Jaccard distance was used to measure the dissimilarity between clusters. Thus, clusters with Jaccard scores closer to 0 represent features which often employed in the same scams, while Jaccard scores closer to 1 represent features which are rarely employed alongside other features.

$$Jaccard's\ Distance = \frac{(A \cup B) - (A \cap B)}{A \cup B}$$

The hierarchical clustering reveals valuable insights into the combinations of social influence features scammers employ. Notably, scarcity and authority are closely clustered, suggesting their concurrent use by cybercriminals. In contrast, reciprocity and consistency clustered with high Jaccard distances. It is likely that the low number of scams using these tactics greatly penalized their clustering. More comprehensive datasets may see more prevalent use of these features and generate more useful insights into their relationships with other social influence features.

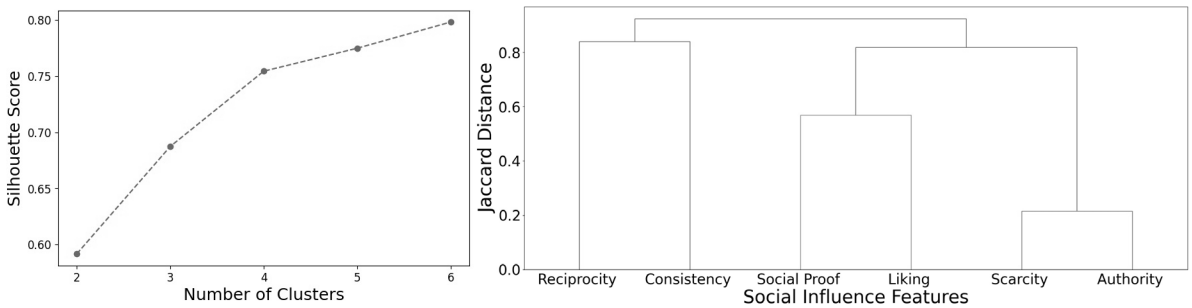


Figure 8: Agglomerative Clustering of Social Influence Features

Figure 9 shows a t-distributed Stochastic Neighbour Embedding (t-SNE) map, which is a valuable tool for mining insights from highly dimensional datasets. Here, t-SNE has been used to reduce the dimensions for the presence and absence of all six social influence tactics into two dimensions. Due to the binary nature of the

encoded data, there are many overlapping datapoints, represented by darker icons. This visualization reveals two key insights into the relationship between social influence tactics and scam types. Firstly, scam types form two distinct clusters, suggesting that the scam type informs which social influence tactics cybercriminals will employ. Secondly, phishing scams use a more complex mixture of social influence tactics compared to fake social media. This is shown by the tight clustering shown by fake social media scams compared to a widely spread distribution for phishing scams.

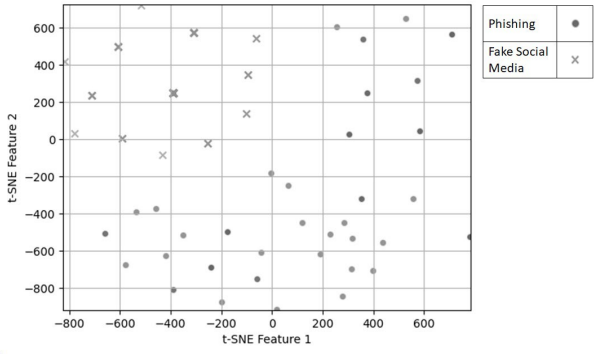


Figure 9: Dimensionality Reduction of Social Influence Features

4.3 Theoretical, Practical and Methodological Implications

Similar to prior studies on cybercrime, using the social influence concepts principals as guiding features for data analysis revealed various insights into psychological tactics employed by cybercriminals from a noisy cybersecurity dataset (Ferreira et al., 2015; Wright et al., 2014). Our study also suggests that integrating features developed by behavioural scientists and techniques from data science can produce high quality and usable insights from noisy scam datasets (Karpatne et al., 2017). In our case, selecting features from the social influence model enriched and improved the value and interpretability of our cybercrime intelligence dataset. For instance, the close clustering of scarcity and authority we found suggests that cybercriminals have a sophisticated understanding of human psychology and are seeking to both exploit the victim's perceived rarity of opportunities (scarcity) and manipulate their perceived trust in certain institutions and personalities (authority). Moreover, cybercriminals employing SNS scams focus their deception efforts on appearing familiar and similar to potential victims so that they appear to be friends (liking). These insights can be used to bolster the security posture of organizations. For instance, our theory-guided feature selection can play a crucial role in offender behaviour analytics (OBA) by enabling analysts to predict the type of cybercriminal social engineering activities from emerging threat patterns that can better inform cyberthreat response (Bhowmik and Bhowmik, 2022). For instance, by prioritising the most prevalent features at a point in time, cybersecurity experts can develop more robust techniques to identify unusual activities and initiate timely incident responses. Moreover, they can use the insights from feature selection to shape cybersecurity awareness training programs. By identifying common attack vectors and common tactics employed by cybercriminals, training can become more targeted and directly applicable to the most recent cybersecurity challenges faced by individuals and organizations. For example, our cluster analysis findings illustrated above suggest that awareness initiatives should highlight the interplay between scarcity and authority, by emphasising common scenarios where these tactics are employed. Educating users about these dual tactics can empower them to recognize and resist these types of manipulative social engineering schemes. We also introduced a robust collaboration architecture that outlines a systematic approach to theory-guided feature selection in the analysis of cybercrime datasets. This framework integrates insights from domain experts and data scientists, ensuring a holistic exploration of internal and external cybercrime feeds (Mao et al., 2019). Furthermore, we also propose a phased approach to feature selected data analysis, which includes preparation of the feature-selected dataset, exploratory data analysis where key patterns and insights are uncovered, and although not within the scope of this study, the application of automated selection techniques for model training and evaluation (Hastie et al., 2001). As found in other disciplines, we believe that theory-guided feature selection will contribute to the accuracy and effectiveness of machine learning models and that models trained on theory-guided features will generally outperform generic data science models for complex cyber intelligence datasets (Faghmous and Kumar, 2014; Karpatne et al., 2017). Finally, our findings underscore the need for multifaceted countermeasures in cybersecurity that focus not only on technical defences but also incorporate psychological and behavioural approaches.

5. Conclusions and Future Work

In this pilot study, we explored the feasibility of theory-guided feature selection in cybercrime data science. One of the challenges of cybercrime intelligence datafeeds and other vast and complex big data sets used in cybersecurity intelligence and analytics is the high severity of 'noisy' features. Our investigation suggests that integrating behavioural science theories to identify and emphasize features that are theoretically relevant to the problem at hand can reduce the noise levels in cybercrime data science models. Towards this end, we applied the psychological and behavioural insights of social influence theory to enrich and improve the value and interpretability of a dataset of scams involving complex social engineering techniques (Cialdini and Goldstein, 2004; Ferreira et al., 2015). Despite the small scale of our analysis, the improved interpretability of our dataset suggests that the proposed methodology could be further explored and refined in future research with larger datasets. We observed the noticeable relevance of selected features such as liking and social proof and their relationship with scam types, such as phishing and fake social media scams. Cybersecurity practitioners can apply the theory-guided feature selection approach to gain insights about cybercrime patterns and apply these to enhance the effectiveness of cyberthreat response and cybersecurity awareness training (Korpela, 2015). In offender behaviour analytics (OBA), feature selection methods can also help identify the social engineering modus operandi that are relevant to cybercrime victimization for different scam types. Our approach also underscores the importance of fostering collaboration between behavioural scientists and data scientists in cybersecurity data science teams to enhance data-driven decision-making processes. This study builds upon prior research by emphasizing the need for theory-guided cybersecurity data science. While the existing cybersecurity literature acknowledges the importance of feature selection, our contribution lies in the systematic integration of social influence theory to scam data involving social engineering techniques to enhance the overall interpretability and predictive power of cybercrime data science models. One of the main limitations of our pilot study is the notably small size and scope of our dataset which limits the generalizability of our results. Another limitation is that our study is also dependent on our expert understanding of social influence theory which is also evolving. Moreover, the binary encoding of features restricted the types of data science models that we could apply. Future research could use scales with increased granularity to encode the social influence tactics in scams so that the data could be analysed and visualized using more robust techniques such as linear regression and PCA, to provide more detailed and nuanced insights. These limitations highlight the need for continued refinement and adaptation of our social influence feature selection methodology. In future research, we intend to develop an automated pipeline for identifying and encoding social influence features from cybercrime intelligence datafeeds and scam images. We will refine and train our model on larger datasets, using automated feature selection algorithms and test the ranking and importance of our features as well as model performance (Raschka et al., 2020). Our model will also be expanded to use a wider range of machine learning tools and evaluation metrics to extract more meaningful insights from cybercrime intelligence datafeeds. We call for the further development of theory-guided data science methods that harnesses the contextual understanding and explanatory power of the behavioural sciences with the predictive power of data science to improve decision-making in cybersecurity. We hope that this pilot study will serve as a stepping stone for the further exploration of theory-guided data science and theory-guided feature selection to fruitfully harvest the ever-expanding, increasingly complex and noisy cybercrime and cybersecurity datasets.

Acknowledgements

This work has been supported by the National Research Foundation (NRF).

References

Bechor, T., Jung, B., 2019. Current State and Modeling of Research Topics in Cybersecurity and Data Science. Systemics, Cybernetics and Informatics 17, 129–156.

Bhowmik, P., Bhowmik, P.C., 2022. A Machine Learning Approach for Phishing Websites Prediction with Novel Feature Selection Framework, in: Hossain, S., Hossain, Md.S., Kaiser, M.S., Majumder, S.P., Ray, K. (Eds.), Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021, Lecture Notes in Networks and Systems. Springer Nature, Singapore, pp. 357–370.

Chen, Chiang, Storey, 2012. Business Intelligence and Analytics: From Big Data to Big Impact. MIS Quarterly 36, 1165–1188.

Cialdini, R.B., Goldstein, N.J., 2004. Social Influence: Compliance and Conformity. Annual Review of Psychology 55, 591–621.

Downton, J.E., Collet, O., Hampson, D.P., Colwell, T., 2020. Theory-guided data science-based reservoir prediction of a North Sea oil field. The Leading Edge 39, 742–750.

Embarak, Dr.O., 2018. Data Analysis and Visualization Using Python: Analyze Data to Create Visualizations for BI Systems. Apress, Berkeley, CA.

Faghmous, J.H., Kumar, V., 2014. A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science. Big Data 2, 155–163.

Ferreira, A., Coventry, L., Lenzini, G., 2015. Principles of Persuasion in Social Engineering and Their Use in Phishing, in: Tryfonas, T., Askoxylakis, I. (Eds.), Human Aspects of Information Security, Privacy, and Trust, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 36–47.

Foroughi, F., Luksch, P., 2018. Data Science Methodology for Cybersecurity Projects, in: Computer Science & Information Technology. Presented at the 5th International Conference on Artificial Intelligence and Applications, Academy & Industry Research Collaboration Center (AIRCC), pp. 01–14.

Gupta, S., Gupta, A., 2019. Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. Procedia Computer Science 161, 466–474.

Hastie, T., Friedman, J., Tibshirani, R., 2001. The Elements of Statistical Learning, Springer Series in Statistics. Springer, New York, NY.

Heartfield, R., Loukas, G., Gan, D., 2016. You Are Probably Not the Weakest Link: Towards Practical Prediction of Susceptibility to Semantic Social Engineering Attacks. IEEE Access 4, 6910–6928.

Kamenov, D., 2018. Intelligent Methods for Big Data Analytics and Cyber Security. ISIJ 39, 255–262. https://doi.org/10.11610/isij.3921

Karpatne, A., Atluri, G., Faghmous, J.H., Steinbach, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331.

Korpela, K., 2015. Improving Cyber Security Awareness and Training Programs with Data Analytics. Information Security Journal: A Global Perspective 24, 72–77.

Li, C., Xie, W., Zhou, K., 2018. Efficient Binary-Encoding Access Control Policy Combination for Large-Scale Collaborative Scenarios, in: 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE). Presented at the 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 560–566.

Mao, Y., Wang, D., Muller, M., Varshney, K.R., Baldini, I., Dugan, C., Mojsilović, A., 2019. How Data ScientistsWork Together With Domain Experts in Scientific Collaborations: To Find The Right Answer Or To Ask The Right Question? Proc. ACM Hum.-Comput. Interact. 3, 1–23.

McAfee, A., Brynjolfsson, E., Davenport, T.H., Patil, D.J., Barton, D., 2012. Big Data: The Management Revolution. Harvard business review 90, 60–68.

Miao, J., Niu, L., 2016. A Survey on Feature Selection. Procedia Computer Science 91, 919–926.

Milgram, S., 1963. Behavioral Study of obedience. The Journal of Abnormal and Social Psychology 67, 371–378.

Naidoo, R., 2020. A multi-level influence model of COVID-19 themed cybercrime. European Journal of Information Systems 29, 306–321.

Raschka, S., Patterson, J., Nolet, C., 2020. Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. Information 11, 193.

Rassam, M.A., Maarof, M.A., Zainal, A., 2017. Big Data Analytics Adoption for Cyber- security: A Review of Current Solutions, Requirements, Challenges and Trends. Journal of Information Assurance and Security. 11, 124–145.

Rawat, D.B., Doku, R., Garuba, M., 2021. Cybersecurity in Big Data Era: From Securing Big Data to Data-Driven Security. IEEE Trans. Serv. Comput. 14, 2055–2072.

Sarker, I.H., Kayes, A.S.M., Badsha, S., Alqahtani, H., Watters, P., Ng, A., 2020. Cybersecurity data science: an overview from machine learning perspective. J Big Data 7, 41.

Slawski, M., Zu Castell, W., Tutz, G., 2010. Feature selection guided by structural information. Ann. Appl. Stat. 4.

Stajano, F., Wilson, P., 2011. Understanding scam victims: seven principles for systems security. Commun. ACM 54, 70–75.

Suraj, M.V., Kumar Singh, N., Tomar, D.S., 2018. Big data Analytics of cyber attacks: a review, in: 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCA). Presented at the 2018 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN), IEEE, Pondicherry, pp. 1–7.

Viaene, S., 2013. Data Scientists Aren't Domain Experts. IT Prof. 15, 12–17.

Wagner, N., Rondinelli, J.M., 2016. Theory-Guided Machine Learning in Materials Science. Front. Mater. 3, 1–9. https://doi.org/10.3389/fmats.2016.00028

Wang, L., Jones, R., 2021. Big Data Analytics in Cyber Security: Network Traffic and Attacks. Journal of Computer Information Systems 61, 410–417.

Wright, R.T., Jensen, M.L., Thatcher, J.B., Dinger, M., Marett, K., 2014. Research Note—Influence Techniques in Phishing Attacks: An Examination of Vulnerability and Resistance. Information Systems Research 25, 385–400.

Zhu, X., Wu, X., 2004. Class Noise vs. Attribute Noise: A Quantitative Study. Artificial Intelligence Review 22, 177–210.

Zhu, X., Wu, X., Chen, Q., 2003. Eliminating Class Noise in Large Datasets, in: Proceedings of the Twentieth International Conference on Machine Learning. Presented at the ICML-2003, Washington DC, pp. 1–18.

Infusing Morabaraba Game Design to Develop a Cybersecurity Awareness Game (CyberMoraba)

Mike Nkongolo
University of Pretoria, Faculty of Informatics, South Africa

Mike.wankongolo@up.ac.za

Abstract: Numerous studies have confirmed the effectiveness of Cybersecurity Awareness Games (CAGs) in enhancing the security posture of diverse organizations. As these organizations increasingly face the formidable challenge of cyberattacks, implementing serious CAGs to solve this issue has become a paramount concern. This article introduces an innovative approach to cybersecurity education by presenting a serious CAG. The game aims to effectively educate students about critical aspects of cybersecurity awareness engagingly and interactively. The study aimed to redefine cybersecurity awareness training by introducing an indigenous game design that intricately incorporates the traditional South African Morabaraba board game. While the effectiveness of non-indigenous games like "Capture The Flag (CTF)" in cybersecurity training is acknowledged, indigenous designs have been overlooked. This research creatively integrates Morabaraba's gameplay into cybersecurity training, adapting it into a competitive game where players adopt the roles of either defenders or attackers, with corresponding tokens/images symbolizing various cyber defense and attack strategies. Both the defenders and attackers in the game can elevate their awareness scores by strategically positioning defensive or attacking images on the game board. Subsequently, a judging entity assesses the players' moves and assigns scores based on the accuracy of the images placed. The game mirrors real-world scenarios, promoting strategic thinking and leveraging interactive gameplay for practical insights into cybersecurity awareness. Players demonstrate their cybersecurity knowledge through offensive and defensive strategies. A group of 40 students evaluated the game's effectiveness, highlighting its potential to create an engaging and competitive learning environment that imparts cybersecurity principles and practical application. The evaluation of the game mechanics demonstrated a remarkably positive outcome, with students expressing both enjoyment and an enhanced understanding of cybersecurity awareness.

Keywords: Cybersecurity, Serious game, Game design, Game mechanics, Cybersecurity awareness, Game theory, Education

1. Introduction

Cybersecurity awareness refers to the knowledge, understanding, and vigilance that individuals, organizations, and communities possess regarding potential cybersecurity threats, risks, and best practices (Abd et al, 2015). It involves being informed about various cyber threats, understanding how to protect sensitive information, and adopting proactive measures to prevent cyberattacks. Cybersecurity awareness is vital for safeguarding individuals and organizations from a wide range of cyber threats, including viruses, malware, phishing, and ransomware (Chowdhury & Gkioulos, 2021; Abd et al, 2015). It aids in preventing data breaches, financial losses, and reputational damage, while also educating about personal privacy protection and secure online practices. Moreover, it ensures compliance with industry regulations, defends against social engineering tactics, and safeguards critical infrastructure from potential disruptions (Naidoo & Jacobs, 2023; Nkongolo & Tokmak, 2023). Embracing cybersecurity awareness reflects a global responsibility to maintain a secure digital landscape, fostering responsible online behavior and contributing to a safer internet for everyone. The impact of comprehensive cybersecurity awareness training and education is profound. When individuals are well-trained and educated about cybersecurity, their ability to recognize, prevent, and respond to cyber threats significantly improves. This leads to a strengthened overall security posture for both individuals and organizations (Abd et al, 2015). Informed individuals are better equipped to identify phishing attempts, avoid malicious links, protect sensitive information, and practice secure online behaviors (Chowdhury & Gkioulos, 2021). Moreover, they contribute to creating a culture of cybersecurity, where knowledge is shared and best practices are followed, enhancing the collective resilience against cyberattacks. Ultimately, investing in cybersecurity awareness training results in a more vigilant, proactive, and security-conscious digital community (Chowdhury & Gkioulos, 2021; Naidoo & Jacobs, 2023). In this paper, we present "CyberMoraba," a game meticulously crafted to educate students about the nuances of cybersecurity awareness. Numerous games have been suggested to aid in combating cybercrime. Prominent games encompass CyberCIEGE, which adeptly simulates real-world cyber incidents and furnishes players with a platform to hone their defense against cyber threats (Irvine et al, 2005). Similarly, CTRL-ALT-Hack immerses players in simulated hacking scenarios, fostering a hacker's perspective to bolster defenses against cyber threats (Kipker et al, 2021). CyberStart ingeniously imparts cybersecurity skills to students, propelling them to conquer challenges and puzzles (Stoker et al, 2021), while R00tz Asylum intricately weaves interactive games into the domain of young individuals, imparting knowledge about hacking, coding, and cybersecurity (Haroldson & Ballard, 2021). Further, Cybersecurity Capture The Flag (CTF) stands as an engaging challenge where participants decode cybersecurity puzzles and execute tasks to amass points (Ortiz-

Garces et al, 2023). Nonetheless, the distinguishing aspect among these games lies in their omission of indigenous game design. This observation underscores the untapped potential within the realm of cybersecurity awareness training. The incorporation of indigenous game design offers a distinct avenue for imparting knowledge and understanding, aligning with cultural sensibilities and traditional narratives. By embracing indigenous game elements, a unique bridge is formed between cybersecurity education and heritage, fostering a more resonant and relatable approach to raising awareness. This symbiosis not only introduces fresh perspectives to cybersecurity but also enriches the fabric of educational experiences by harnessing the power of cultural familiarity. To the best of our knowledge, there has been no instance of a serious CAG being developed with an indigenous gameplay such as Morabaraba. Morabaraba, also known as "Umlabalaba," is a traditional two-player strategy board game with African origins (Nkopodi & Mosimege, 2009; Nkongolo, 2023), particularly prevalent among the Bantu-speaking peoples of South Africa. The game is played on a distinctive board with three concentric squares, each comprised of eight intersection points (Nkopodi & Mosimege, 2009; Nkongolo, 2023). Players have twelve pieces, typically represented by cowrie shells or stones, and the objective is to strategically move these pieces to form specific patterns and capture the opponent's pieces (Nkopodi & Mosimege, 2009; Nkongolo, 2023). Morabaraba is known for its intricate gameplay, requiring careful planning, tactical thinking, and adaptability (Nkopodi & Mosimege, 2009; Nkongolo, 2023). It has cultural and social significance within various African communities, serving as a means of entertainment and often symbolizing intelligence and strategic prowess (Nkongolo, 2023). It is a traditional two-player strategy board game that holds cultural significance in the region (Nkopodi & Mosimege, 2009; Nkongolo, 2023). The game is characterized by a distinctive game board layout and involves players attempting to align their game pieces to form rows, thereby gaining an advantage over their opponent (Nkongolo, 2023). It has historical roots within indigenous communities and continues to be played in both traditional and modern settings (Nkopodi & Mosimege, 2009). The Morabaraba game's strategic elements offer innovative avenues to elevate cybersecurity awareness training. The gameplay of Morabaraba is intricate and strategic. Two players, each with twelve pieces, take turns to make their moves on the board, aiming to create specific patterns and capture their opponent's pieces (Nkopodi & Mosimege, 2009). The game progresses in two phases:

Placement phase: In this initial phase, players take turns placing their twelve pieces on the board's intersection points, one piece at a time (Nkopodi & Mosimege, 2009). The goal is to strategically position the pieces for future movement and pattern creation.

Movement phase: Once all pieces are placed, the game transitions into the movement phase. During this phase, players take turns moving one of their pieces along the lines on the board, attempting to form specific patterns (Nkopodi & Mosimege, 2009). The patterns include creating a "Mill," which consists of three of their pieces in a straight line, horizontally or vertically (Nkopodi & Mosimege, 2009). When a player forms a Mill, they are allowed to capture one of their opponent's pieces. The game continues until one player can no longer make a legal move or is reduced to only three pieces (Nkopodi & Mosimege, 2009), which allows the opponent to capture their pieces freely. The player who successfully captures all of their opponent's pieces or immobilizes them wins the game. Morabaraba's gameplay requires careful planning, strategic thinking, and adaptability, making it a challenging and engaging traditional board game (Nkopodi & Mosimege, 2009; Nkongolo, 2023). Integrating Morabaraba's strategic dynamics into cybersecurity education not only enhances participants' understanding but also makes training engaging and relatable. This approach effectively combines traditional game principles with modern security challenges, fostering practical skills acquisition and bridging the gap between entertainment and education in the realm of cybersecurity. The research question is formulated as follows: "How does the integration of traditional South African Morabaraba gameplay into cybersecurity training, as exemplified by CyberMoraba, impact the effectiveness of cybersecurity awareness and practical application among students?"

This research question explores the influence of introducing Morabaraba-inspired elements into cybersecurity training, using CyberMoraba as a model. It focuses on assessing how this unique approach affects students' cybersecurity awareness and their ability to practically apply cybersecurity principles. The question aims to measure the educational value and effectiveness of such an indigenous game design in enhancing cybersecurity knowledge and skills. The target group for the CyberMoraba game is a critical aspect of its design and implementation. It is not a one-size-fits-all approach, as different groups within an organization may have varying levels of cybersecurity awareness and different roles and responsibilities in ensuring cybersecurity (Abd et al, 2015; Ortiz-Garces et al, 2023; Naidoo & Jacobs, 2023). For this study, the primary target audience is the students, and it is essential to customize the game's usage based on their specific needs. However, within an organizational context, various personnel groups, such as HR (Human Resources), finance, and technical

personnel, may have unique cybersecurity challenges and requirements, and the game can be adapted to address these distinct aspects (Ortiz-Garces et al, 2023). In essence, the game's flexibility and versatility are key strengths, allowing organizations to tailor its use to specific target groups, and ensuring that it remains relevant and impactful across various departments and individuals within the organization.

2. Methodology

In this section, we illuminate various facets of the proposed game's design. We commence by outlining the formulation of the game's rules, subsequently delving into an elucidation of the constituents that constitute the game environment, serving as the backdrop for the gameplay itself. The discourse then concludes by elaborating upon the game's scoring mechanics and the technological architecture underpinning the Graphical User Interface (GUI). The components that comprise the game GUI are visually depicted in Figure 1.

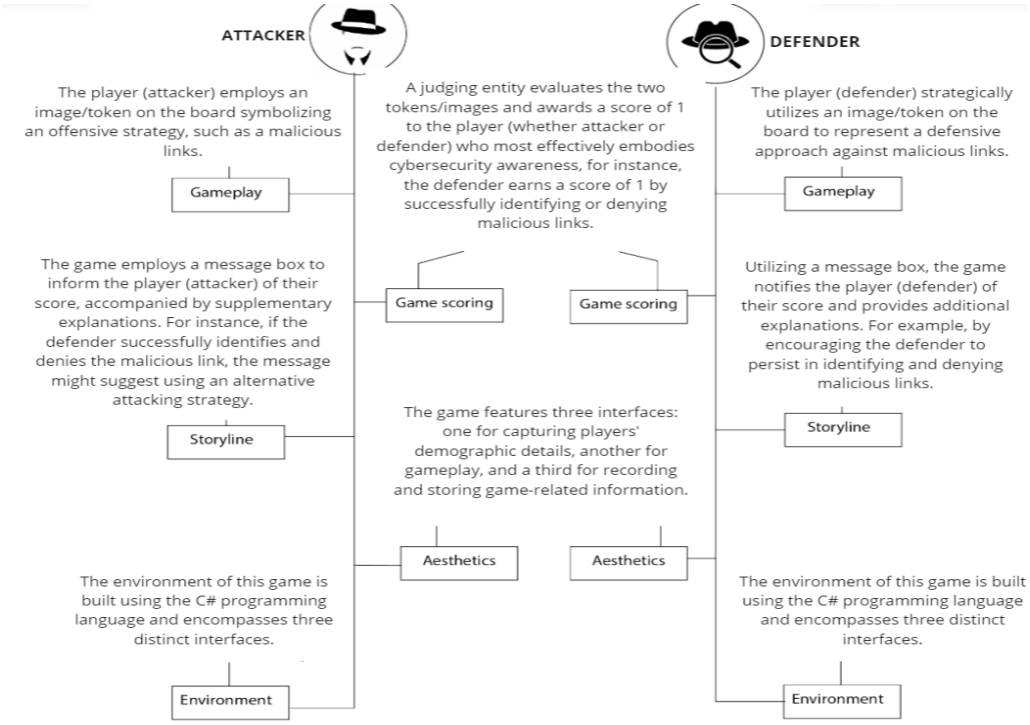


Figure 1: The game components

2.1 Rules of the Game

Each player is given a set of 13 tokens or images that represent attacking or defending strategies (Table 1). The attacker's tokens are labelled with A1, A2, A3, ..., A13, and the defender's tokens are labelled with D1, D2, D3, ..., D13 (Table 1). During the game, each player takes turns placing their tokens on the game board, with each token placement representing a move that determines whether the attacker or defender used an optimal strategy (Nkongolo, 2023). Commencing with an unoccupied board, the game's initiation rests with the attacker. Subsequently, the judging entity bestows a positive reward of 1 upon the player executing the most adept move, while assigning a zero to the player opting for an incorrect move (Traulsen & Glynatsi, 2023). Progress within the game materializes through the placement of the appropriate tokens onto the board. The game format is designed to simulate real-world cybersecurity scenarios and help a person gain a deeper understanding of cybersecurity best practices (Abd et al, 2015). The game's clear rules and sequential nature make it easy for players to understand and engage with, while the judging entity motivates players to perform their best. The strategies of the game require players to think strategically about their moves and anticipate their opponent's next move (Nkongolo, 2023). Successful players need to balance their offensive and defensive moves to gain an advantage on the board. The objective of the game is to simulate a scenario in which the defender and attacker compete to accomplish their respective objectives (Traulsen & Glynatsi, 2023).

The defender aims to obtain the highest cybersecurity awareness score by either correctly placing the appropriate image on the board or using the most efficient cybersecurity strategy to defend against the attacker's actions. Conversely, the attacker strives to achieve the highest intrusion score by implementing the most effective attacking strategy to penetrate the defender's cybersecurity defense.

Table 1: The attacker and defender tokens

Token	Image	Definition	Token	Image	Definition
A1	Email	Malicious e-mail	D1	Denying	Blocking/denying actions
A2	Phone	Malicious phone call	D2	Network monitoring	Network traffic analysis
A3	Chat	Malicious chat	D3	Avoid clicking	Refuse to click
A4	Attachment	Malicious attachment	D4	Identification	Verification process
A5	Donate	Malicious directory	D5	No trust	Zero trust policy
A6	Password	Malicious directory	D6	Upload	Uploading process
A7	Connection	Malicious network connection	D7	Trust	The defender trusts
A8	Access	Malicious intrusion	D8	Provide	Providing information
A9	Data	Malicious data	D9	Confidential	Confidentiality of data
A10	Data loss	Data loss process	D10	Report	Reporting cyber incidents
A11	Click	Malicious link	D11	Social media	Sharing data on social media
A12	Sensitive data	Theft of data	D12	Connection	Secured network connection
A13	Message	Malicious communication	D13	Backup	Data recovery

Winning the game necessitates that the defender understands the various cybernetic strategies that the attacker may utilize and can select the optimal defense approach to combat each one. The attacker must leverage their knowledge of data protection vulnerabilities to use the correct image to score points. In the same vein, the defender must be watchful and put the correct image on the board to thwart the attacker's successful attacks. To win the game, it is essential to have a comprehensive data protection strategy that includes both technical and non-technical measures. To compute the final score of each player when the game is over, the judging entity evaluates the tokens as illustrated in Table 2 where A% and D% represent the attacker and defender score (Traulsen & Glynatsi, 2023; Nkongolo, 2023). It is important to note that the types and number of images can be expanded. Figure 1 and Figure 2 present the attacker and defender tokens/images (Figure 3). The game comes to an end when all 13 images have been placed on the board. If both players have the same total score, the game is drawn.

2.2 Architecture of the Game

The game's software was developed using Visual Studio Integrated Development Environment (IDE) and necessitates a desktop or personal computer equipped with a connected mouse and keyboard to facilitate gameplay (Traulsen & Glynatsi, 2023). Elevating the game's challenge can be achieved by incorporating a timer mechanism. In the game's sequence of play, the player initiates by submitting demographic details and choosing their preferred role, either attacker or defender (Figure 4). Subsequently, they access the game board through a menu, where they select images to signify their strategic moves (Figure 5 and Figure 6).

Table 2: An instance of the judging scoring scheme

	A	D	A%	D%	Judge	Message
1	Email	Zero trust	0	1	Defender best move	Never trust malicious emails
2	Click	Denying	0	1	Defender best move	Keep denying malicious links
3	Chat	Identification	0	1	Defender best move	Identification of malicious chats
4	Phone call	Trust	1	0	Attacker best move	The defender trusted a malicious call
5	Connection	Connection	0	1	Defender best move	Secured connection suggested
6	Access	Identification	0	1	Defender best move	Malicious access identified
7	Data loss	Upload	1	0	Attacker best move	Data loss occurred
8	Click	Provide	1	0	Attacker best move	Malicious link used
	Total		3%	5%	Defender wins with 5%	Defender best moves: zero trust, denying, identification, secured connection

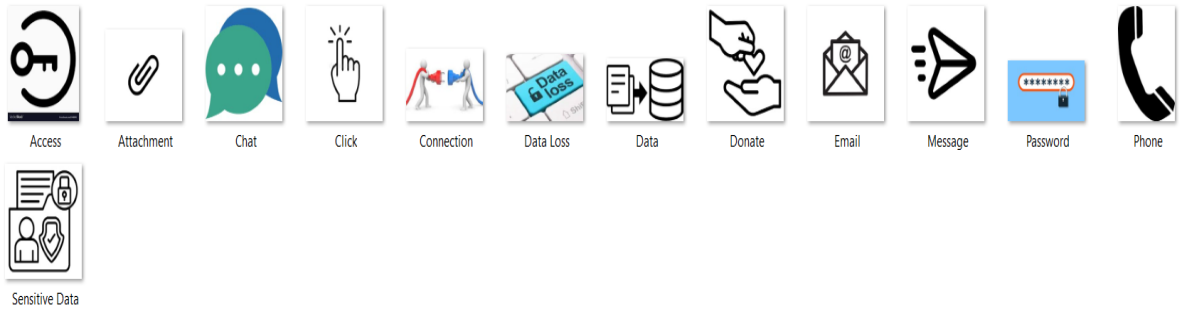


Figure 2: The attacker tokens (author's images)

In the left depiction of Figure 6, the attacker successfully convinces the defender to access a malicious website, leading to the compromise of the defender's credentials. As a result, the attacker achieves a score of 1. In the right depiction of Figure 6, the attacker proposes a deceitful connection, met with the defender's skepticism. The defender recommends a secure alternative, showcasing resilience, and earns a score of 1. Once all images have been utilized by both players, the game concludes, and the record menu will save players' information (Figure 7). The scoring mechanism employed by the judging entity is implemented with a conditional "if-else" structure based on Table 2.



Figure 3: The defender tokens (author's images)

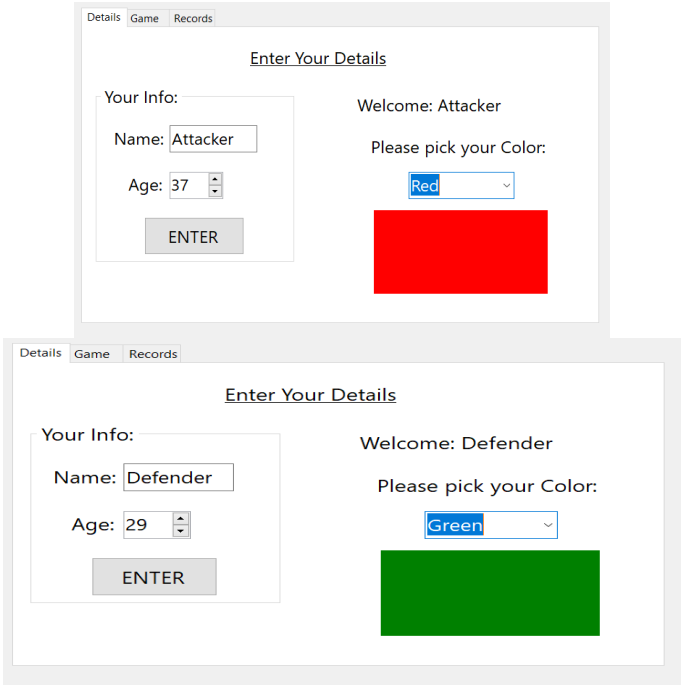


Figure 4: The player role selection



Figure 5: Launching the gameplay scenario

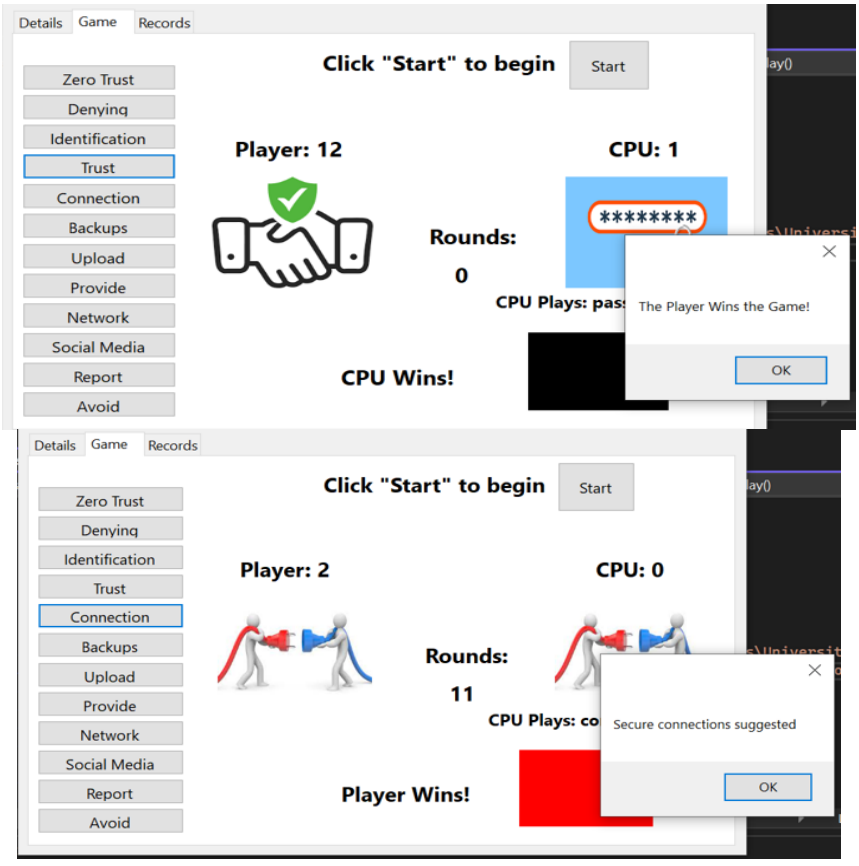


Figure 6: Illustration of a scenario involving game mechanics

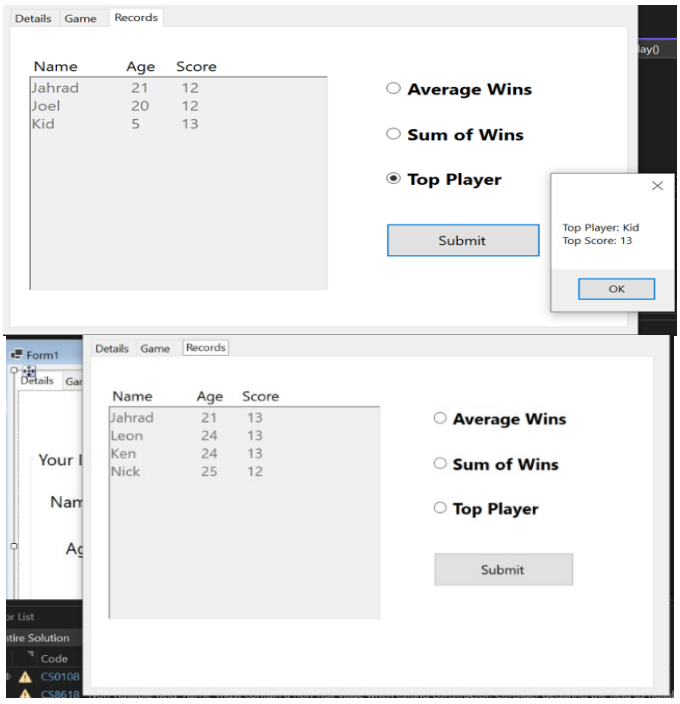


Figure 7: Illustration of recorded players' information

3. Results of Players' Feedback Assessment

For this research, a cohort of 40 student participants was meticulously chosen from the university community, encompassing diverse backgrounds and possessing a commendable familiarity with gaming dynamics as well as a solid understanding of the realm of cybersecurity awareness. Participants were not allowed to specify improvement areas in the questions provided. Additionally, questions related to the indigenous game design element were not included in the questionnaire. The survey administered to the participants encompassed the following inquiries:

On a scale of 0 to 10, with 0 indicating unlikeliness and 10 signifying utmost likelihood, how probable would you be to engage in this gameplay?

This question seeks to understand students' level of interest and willingness to engage in gameplay. The findings from this question can provide insights into the overall appeal of the game. If a majority of students rate their likelihood to engage at higher values (7-10), it suggests that the gameplay has garnered significant interest (Figure 8 and Figure 9).

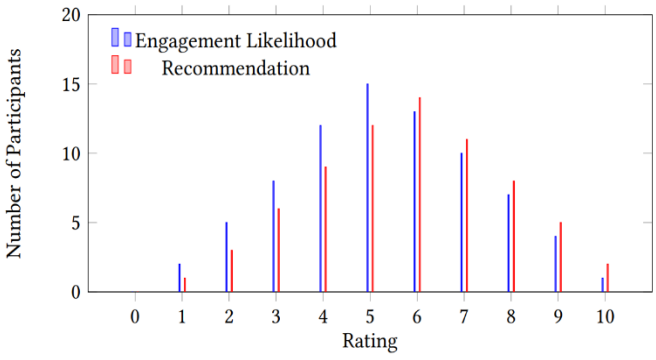


Figure 8: Participants gameplay interest

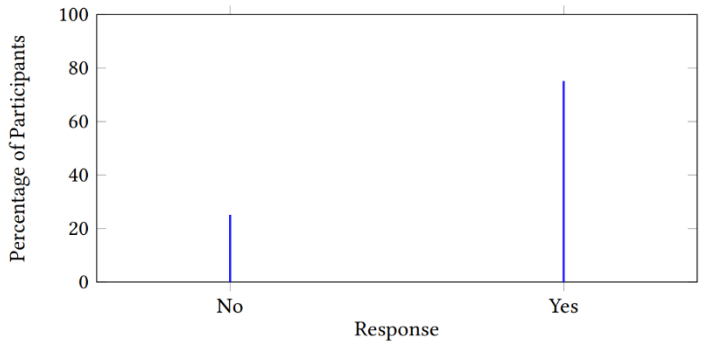


Figure 9: Participants gameplay responses

Conversely, lower ratings (0-6) might indicate areas where the game could be improved to increase engagement (Figure 8).

Would you consider recommending this game to others for their participation?

This question delves into students' willingness to recommend the game to others. Positive responses (Yes) indicate that students found the game enjoyable and valuable enough to recommend to their peers (Figure 10).

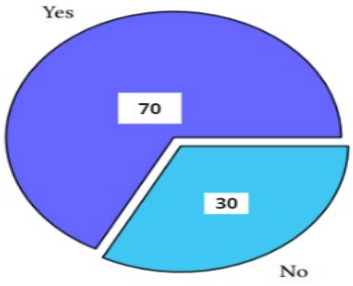


Figure 10: Perceived potential of the game to educate

Conversely, negative responses (No) might imply that students identified areas where the game could be enhanced before they would recommend it.

What particular aspects of this game would you find enjoyable and engaging?

Figure 11 indicates that students found various aspects of the game enjoyable and engaging, with the highest percentage leaning toward gameplay, followed by challenges and interactive elements. This information can help guide future game design decisions and improvements to enhance the overall engagement and effectiveness of the game in promoting cybersecurity awareness.

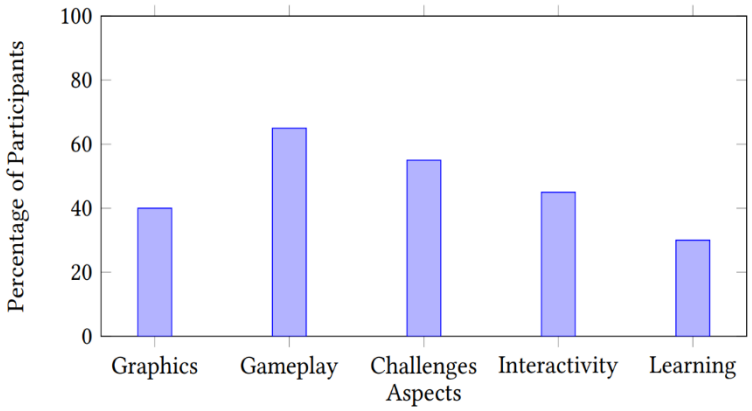


Figure 11: Participants perceptions of various aspects of the game

In your perspective, what enhancements or modifications could be made to elevate the game's overall experience?

Figure 12 demonstrates that students had varied opinions on how to enhance the game's overall experience. The top three categories of suggestions—More Challenges, Improved Interactivity, and Better Graphics—reflect students' desire for engaging gameplay with higher levels of complexity, interactive elements, and visual appeal. This information is valuable for informing game development decisions, aligning with participants' preferences, and crafting a more enjoyable and impactful game for promoting cybersecurity awareness.

4. Discussion and Principal Findings

The emergence of serious games within the realm of cybersecurity is undergoing rapid expansion (Abd et al, 2015). This research delves into the intricate landscape of constructing a potent serious game designed in the context of cybersecurity awareness (Nkongolo, 2023). Through our exploration, we aim to provide readers with a comprehensive insight into harnessing indigenous game designs to fashion efficacious serious games for countering cyber threats (Chowdhury & Gkioulos, 2021). Notably, the players/students' responses indicated a strong engagement with both the gameplay and storyline—a pivotal facet. Achieving equilibrium between entertainment and education is a paramount objective for serious games, ensuring the creation of an environment conducive to effective learning (Calvano et al, 2023).

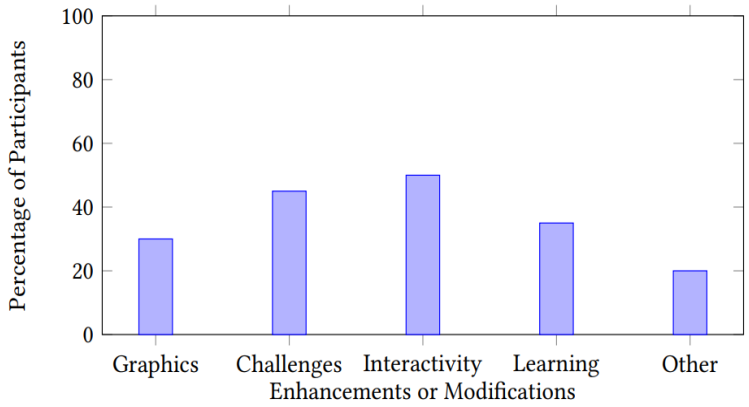


Figure 12: Participants' perceptions of gameplay modification

We elucidated the methodology for crafting a game that seamlessly blends enjoyment and educational value by showcasing functional mechanics. This elucidation carries significance, as a game that fails to resonate positively with players runs the risk of being ineffective as a learning tool (Moumouh et al, 2023). Consequently, the vital synergy between enjoyment and learning stands as a fundamental principle for the efficacy of serious CAGs (Traulsen & Glynatsi, 2023). The research introduces several noteworthy contributions, including innovative game mechanics, a multi-interface gameplay approach, and enhanced gameplay elements. The unique mechanics involve players/students assuming roles as defenders or attackers to address cyber threats, fostering a better understanding of cybersecurity. The game employs three C# interfaces to create a diverse and comprehensive learning environment. Additional features like timers and tokens/images enhance the gameplay, providing more challenges and learning opportunities. Named "CyberMoraba," the game creatively blends traditional Morabaraba gameplay with cybersecurity awareness education. Morabaraba and CyberMoraba share a fundamental similarity in their competitive and strategic gameplay. Both games are designed to engage players in strategic thinking and decision-making, making them ideal platforms for learning and skill development:

Competitive nature: Morabaraba, is known for its competitive gameplay (Nkopodi & Mosimege, 2009). Similarly, CyberMoraba introduces a competitive element to cybersecurity awareness training. In CyberMoraba, players take on the roles of defenders and attackers, competing to outsmart each other in a simulated cyber environment. This competitive aspect adds an element of excitement and engagement to the learning process in both games.

Strategy and decision-making: Morabaraba requires players to think strategically and plan their moves carefully (Nkopodi & Mosimege, 2009). The placement of tokens on the board can significantly impact the outcome of the game. CyberMoraba extends this strategic thinking to the realm of cybersecurity. Players must make informed decisions on where to position defensive and attacking images on the game board to maximize their awareness scores. This strategic element encourages critical thinking and problem-solving, a common trait in both games.

Real-world simulation: Both Morabaraba and CyberMoraba offer a degree of real-world simulation. In Morabaraba, players replicate the strategic thinking used in traditional cattle herding (Nkopodi & Mosimege, 2009). In CyberMoraba, the game mirrors real-world cybersecurity scenarios, allowing players to apply their knowledge practically. This real-world context is a shared feature that enhances the educational value of both games.

Interactive learning: Morabaraba has long been used as a means of interactive learning, teaching players valuable skills and cultural knowledge (Nkopodi & Mosimege, 2009). CyberMoraba extends this tradition by incorporating elements of cybersecurity awareness and education into its gameplay. Both games provide an interactive and engaging learning experience, making them effective tools for imparting knowledge.

Skill development: Morabaraba hones players' skills in strategy, planning, and decision-making (Nkopodi & Mosimege, 2009). Similarly, CyberMoraba enhances players' skills related to cybersecurity awareness, defense, and attack strategies. Both games focus on skill development within their respective domains. The relationship between Morabaraba and CyberMoraba lies in their shared emphasis on competitive gameplay, strategic thinking, real-world simulation, interactive learning, and skill development (Nkopodi & Mosimege, 2009; Nkongolo, 2023). CyberMoraba builds upon the traditional game of Morabaraba, adapting its gameplay to educate and engage students in the field of cybersecurity.

4.1 Further Research Development

Presently, our efforts are focused on the ongoing improvement of CyberMoraba by introducing multiple 3D levels to enhance its appeal among students. This initiative aligns with our commitment to advancing Educational Serious Game Design (ESGD), underscoring our anticipation of continued endeavors in this domain. Furthermore, we envisage extending the evaluation of CyberMoraba to encompass more students, this expansion promises to provide us with a broader spectrum of insights, enabling us to further refine and enhance the game.

4.2 Challenges and Applicability

While the initial CyberMoraba evaluation with students yielded positive results, it is important to acknowledge the limitations of the study. One limitation is that the study primarily focused on students, and the effectiveness of CyberMoraba with other target groups, such as professionals in various organizations, remains unexplored. Generalizing the game's impact beyond the student demographic should be approached with caution. Additionally, the study did not include specific questions related to the indigenous game design element, which could have provided valuable insights into its effectiveness and cultural relevance. This can be explored in future research with a distinct focus group separate from the student population. Future work directions should aim to address these limitations and broaden the game's applicability. A concrete direction for future research involves conducting similar evaluations with diverse target groups, including professionals and individuals from different cultural backgrounds. This will help determine the game's generalizability and effectiveness in various educational contexts. Moreover, further studies should incorporate questions that specifically assess the impact of the indigenous game design on the learning experience.

5. Conclusion

Within the scope of this research paper, we introduced a novel ESGD aimed at imparting cybersecurity awareness to students by employing the innovative Morabaraba game design. The game's architecture enabled players to adopt the roles of either defenders or attackers, with corresponding tokens/images symbolizing various cyber defense and attack strategies. The initial prototype of the game was presented, followed by a comprehensive evaluation process. Student responses underscored the game's potential as a robust learning tool for cybersecurity awareness, as well as their inclination to recommend it to others. The implemented cybersecurity game effectively addresses a significant void within the African academic landscape, as it constitutes the first of its kind not only in Africa but also worldwide. Capitalizing on the growing allure of video games among the digital-native generation, the integration of CyberMoraba as a vehicle for promoting cybersecurity awareness is poised to invigorate student motivation and enthusiasm toward acquiring essential cybersecurity skills. In the future, the discoveries made in this paper will be subject to comparison with the existing body of literature about ESGD.

Acknowledgement

The author expresses gratitude to the University of Pretoria's Faculty of Engineering, Built Environment, and Information Technology for their support in funding this research project through the Doctorate University Capacity Development Program (UCDP) Grant A1F637.

References

Abd Rahim, N.H., Hamid, S., Kiah, M.L.M., Shamshirband, S. and Furnell, S., 2015. A systematic review of approaches to assessing cybersecurity awareness. Kybernetes, 44(4), pp.606-622.

Calvano, M., Caruso, F., Curci, A., Piccinno, A. and Rossano, V., 2023. A Rapid Review on Serious Games for Cybersecurity Education: Are "Serious" and Gaming Aspects Well Balanced? 9th International Symposium on End-User Development, 6-8 June 2023, Cagliari, Italy. CEUR Workshop Proceedings.

Chowdhury, N. and Gkioulos, V., 2021. Cyber security training for critical infrastructure protection: A literature review. Computer Science Review, 40, p.100361.

Haroldson, R. and Ballard, D., 2021. Alignment and representation in computer science: an analysis of picture books and graphic novels for K-8 students. Computer Science Education, 31(1), pp.4-29.

Irvine, C.E., Thompson, M.F. and Allen, K., 2005. CyberCIEGE: gaming for information assurance. IEEE Security & Privacy, 3(3), pp.61-64.

Kipker, D.K. and Pape, S., 2021. Case Study: Checking a Serious Security-Awareness Game for its Legal Adequacy. Datenschutz und Datensicherheit-DuD, 45, pp.310-314.

Moumouh, C., Chkouri, M.Y. and Fernández-Alemán, J.L., 2023. Cybersecurity Awareness Through Serious Games: A Systematic Literature Review. In International Conference on Networking, Intelligent Systems and Security (pp. 190-199). Springer, Cham.

Naidoo, R. and Jacobs, C., 2023, June. Cyber Warfare and Cyber Terrorism Threats Targeting Critical Infrastructure: A HCPS-based Threat Modelling Intelligence Framework. In ECCWS 2023 22nd European Conference on Cyber Warfare and Security. Academic Conferences and Publishing Limited.

Nkongolo, M., 2023, January. Game Theory based Artificial Player for Morabaraba Game. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1210-1218). IEEE.

Nkongolo, M., Tokmak, M. (2023). Zero-Day Threats Detection for Critical Infrastructures. In: Gerber, A., Coetzee, M. (eds) South African Institute of Computer Scientists and Information Technologists. SAICSIT 2023. Communications in Computer and Information Science, vol 1878. Springer, Cham. https://doi.org/10.1007/978-3-031-39652-6_3

Nkopodi, N. and Mosimege, M., 2009. Incorporating the indigenous game of morabaraba in the learning of mathematics. South African Journal of Education, 29(3).

Ortiz-Garces, I., Gutierrez, R., Guerra, D., Sanchez-Viteri, S. and Villegas-Ch, W., 2023. Development of a Platform for Learning Cybersecurity Using Capturing the Flag Competitions. Electronics, 12(7), p.1753.

Stoker, G., Clark, U., Vanajakumari, M. and Wetherill, W., 2021. Building a Cybersecurity Apprenticeship Program: Early-Stage Success and Some Lessons Learned. Information Systems Education Journal, 19(2), pp.35-44.

Traulsen, A. and Glynatsi, N.E., 2023. The future of theoretical evolutionary game theory. Philosophical Transactions of the Royal Society B, 378(1876), p.20210508.

Bahrun Naim's Hacking Manifesto: How a Historical Model of Cyber Threat Mischaracterization Helps us Diffuse a Dead Cyberterrorist Influencer

Tim Pappa
National Intelligence University, Bethesda, Maryland, USA

timothy.s.pappa@niu.odni.gov

Abstract: The late Bahrun Naim is generally considered one of the most recognizable names of historical Indonesian cyber terrorist actors, where he continues to be characterized as a "computer guru" and hacker who supported Indonesian extremist and Islamic State hacking. This paper is the first reexamination of Naim's lengthy e-book manifesto chapters on hacking, finding that most of the tutorial demonstrations on hacking tools readers might believe Naim performed were screen shots of other online content that Naim appeared to have plagiarized. This practitioner's paper suggests that Naim is a historical model of cyber threat mischaracterization, questioning who Naim was signaling with his hacking manifesto and why so many audiences including cyber threat analysts may have been influenced by reports on him. This paper evaluates Naim's *native* and *researched* technical knowledge based on his manifesto, suggesting the perception of Naim's native technical knowledge may have been embellished. This paper offers a conceptual interdisciplinary model for practitioners and researchers evaluating cyberterrorists' technical pedigrees and social environments. This model can also be applied to countering the range of cyber warfare including narrative warfare conducted by cyberterrorists, often revealing shortcomings in the mythos of cyberterrorists' influence and technical skills.

Keywords: Bahrun Naim, Cyberterrorism, Islamic State, Cyber behavioral analysis, Signaling theory

1. Introduction

The late Bahrun Naim is generally considered one of the most recognizable names of historical Indonesian terrorist actors, even outside of Southeast Asian counterterror watchers. Approximately twenty years ago, Naim was going to college and working at an internet café, but he had grown up attending Islamic boarding schools in Central Java, Indonesia, including schools prominently associated with some of the most significant Indonesian extremist leaders. These personalities joined what became Jemaah Islamiyah, an al Qaeda regional Southeast Asia affiliate responsible for the deadliest attacks throughout Indonesia in the early to mid-2000s.

Naim was also particularly close to many of these personalities who were close to Abu Bakar Ba'asyir, a Central Java extremist figure in Naim's hometown who had been exiled and later imprisoned related to an attack in the early 2000s. Ba'asyir was widely believed to be the spiritual head of Jemaah Islamiyah. Ba'asyir like Naim later pledged his allegiance to the Islamic State.

Naim was arrested in approximately 2010 and served about two-and-a-half years in Indonesian prison on charges related to illegally obtaining ammunition. Following his prison term, he affiliated with Islamic State and later appeared online in Syria, with his family. Before he left Indonesia, local authorities only considered him to be a "minor irritant".

Naim had a vocal role in facilitating or encouraging attack planning in Indonesia and the travel of Indonesian fighters to Syria, until he was killed in Syria in approximately 2018. Naim's role with Islamic State and his prior Indonesian extremist affiliations are questionable, considering he held no known leadership position and appeared to facilitate Indonesian travel to Syria largely on his own or with the help of his two wives. While there is some suggestion that Naim may have been trying to occupy a role with Islamic State as the leading Indonesian figure, much of this is speculation.

While there has been considerable attention to Naim's apparent role as one of the most "notorious" promoters of attacks throughout Southeast Asia, there has been limited to no attention to his writing on computer network operations and his background with computers. Naim has been historically characterized by much of the counterterrorism community as a "computer guru", who appeared to demonstrate his skills by operating multiple accounts on several platforms while being pursued in Indonesia and in Syria before his death in June 2018. Gunaratna (2018) has written extensively on Naim and Southeast Asian terrorism groups. Like Gunaratna, many observers refer to Naim's degree in informatics engineering and communication and his early use of cryptocurrency and encrypted messaging applications to facilitate attack planning as a demonstration of his technical knowledge.

This paper has limited perspective on who Naim's intended audience was and only general background on who some of Naim's peers were outside of these virtual communities. Naim was believed to have a "rolodex" of extremist contacts in Indonesia from his association with personalities affiliated with Hizbut Tahrir and Jamaah Ansharut Dauluh (JAD), which were much more regionally focused Indonesian collectives of extremists.

Naim appeared to be involved in some capacity with some of the other pockets of Indonesians in Syria as well, and these personalities also sent money and used social media. The audiences of cybercriminals or hacktivists of some sort in Indonesia that we might presume Naim wrote this e-book for were widely considered to be young but experienced in terms of carding and obfuscation. The more organized collectives or groups of Indonesian actors who conducted computer network attacks were characterized as "transitional", familiar with website defacements and Distributed Denial of Service (DDoS) attacks but unable to conduct scalable attacks against more desirable hardened networks. There were a handful of hacktivist groups supporting Islamic State in some form during this period and after Naim's death, with similar attack or tool capabilities. While this paper is questions who the target audiences were for Naim's manifesto, titled, "*Sebuah Perjalanan, Rahasia*" or "A Secret Journey", this paper is fundamentally exploring Naim's technical experience and capabilities. While there are understandable limitations in what we know about Naim's social experience at that time, identifying a contemporary cyberterrorist's technical flaws or other questionable technical behaviors could provide an opportunity to technically and informationally counter cyberterrorist cyber warfare.

2. Trying to Understand Naim's Signalling to his Audiences

Signaling theory is a foundational theory of communication, explaining how we evaluate how reliable what someone does is or how reliable what someone says is. Donath (2007) has characterized signaling theory as a framework for understanding and developing the communicative function of behavior. Generally, if the observable or presumed costs of signaling or displaying something appears to be more than the benefit of that communication, then the signal is probably reliable.

Donath and Liu (2006) wrote about people who appear to be or want to be 'fashionable' – for someone to be truly fashionable, there would be costs involved, such as time and attention to trends and to attending fashion events and costs associated with purchasing material and outfits. That reliability would probably be demonstrated in how responsive that person is materially or demonstrably to periodic shifts in fashion, which again would require costs. There would be more costs involved than benefits to someone falsely claiming and appearing to be 'fashionable', so that signal or display we see online is probably reliable. Here, Donath is highlighting the need to also consider the receiver of a signal or display.

Signaling theory also models highly contextual relationships among people that can explain why some signals or displays in that culture or group are reliable and others are not. Families with limited money in some cultures may still spend lavishly on a wedding as a "costly display" to demonstrate that they are wealthy enough to provide a wedding that meets norms, even if they will likely be more impoverished because of that wedding. While that example suggests there is significant cost involved, there is arguably more cultural benefit to that family and their son or daughter getting married to appear to be wealthy. What's most important here according to Donath is that the cost must be considered or occur in the appropriate domain, which in this case is a culture where lavish weddings are important "social capital" among families. If the same wedding occurred outside of this culture and among other people, the cost would be different and thus the signal would be different. Alvard (2023) in this context has characterized signaling theory when it involves people as fundamentally a theory of culture more than a theory of communication.

Donath referred to Zahavi (1977) who emphasized that costly signals are only understood in the domain or environment of that cost. Donath explained further that Zahavi referred to this foundational signaling context as the "handicap principle", where he suggested as an example that gazelles could afford to display what appear to be extravagant jumps even in the vicinity of predators because they must be fast enough to take those risks. But outside of that environment between predator and prey, whether gazelles are fast or not matters less in terms of cost. Donath has discussed this concept further, using biker gang tattoos as an example. While a gang member displaying his tattoo likely is a reliable signal of his gang affiliation, there would be significant costs for that presumed gang member if he was not actually a member of that biker gang and a real member of that gang discovered he was impersonating a gang member. We could imagine other possibilities as well, such as the display of a gang tattoo communicating to a rival gang member that if he chooses to harm that rival gang member, the victim's gang will likely respond – so the display or signal of the tattoo protects both bikers. Donath (2023) and Bliege et al. (2023) have both highlighted the importance of clear displays so that the sender and the receiver understand each other, often for the sake of each other's safety.

Lang et al. (2023) found in a study of commitment in groups that group members who demonstrated the most aggressive commitment to the group at first attracted more potential group members, including people who hoped for benefits without the same level of commitment. But those group members eventually left, as the committed group leader often influenced other group members to likewise demonstrate a similar level of commitment. The most committed groups with the most committed group leaders were found to be the most competitive and often the most successful in this study. Lang suggested that in situations that perhaps demanded "absolute commitment", observationally costly signals or displays did give those groups an advantage over other groups because of the shared commitment in the group. Even if members were forced to commit further, Lang wrote, generally their overall level of commitment was reinforced.

Donath emphasized that communicators want to be perceived *advantageously*, so they will use signals to communicate the cues they want the audience to see. Could Naim's release of this e-book represent a signal or display in his context of "absolute commitment", which might mean that once he released this manifesto that included diagrams and instructions for making bombs, then surely Western intelligence services would focus their efforts on Naim because he would appear to be a rising Indonesian leader in Syria, even if Islamic State leadership generally disregarded him in Syria? Did Naim begin his e-book with a preamble on the second page that appeared to be a disclaimer of sorts regarding his affiliation with the Islamic State, because he was not recognized by Islamic State leadership?

Naim wrote in this preamble that this e- State. However, later in Naim's introductory sections, he wrote that he is not a leader, and he has no desire to be a leader. Naim continued, asking rhetorically if the reader was "astonished" that he would write that he has no desire to be a leader. Naim wrote that he is already in the Islamic book are his "private" writings and that this work does not reflect the writings or works of the Islamic State. This paper questions what Naim was trying to signal or display to several audiences, asking whether Naim rushed this e-book to display an advantageous move among Islamic State leadership and other Indonesian competitors in Syria, for example.

Potz (2023) explored voluntary "costly signals" in religious communities, where he suggested it may have been assumed people who perform costly signals in a religious community are required to accomplish these rituals to be accepted into the group. There are other motivations driving costly signals, however, that are not required by the group, but nonetheless are regarded as costly and valuable, he wrote. Potz characterized these voluntary signals or displays as "structurally empowering" because of how it might influence the group's perceptions of whoever is signaling these costs. While required "costly" signals communicate commitment to that religious community, displaying or signaling a voluntary cost could improve or raise someone's reputation or standing within that community. Potz referred to Toland (1988) who wrote about this legitimation of power as something to be "thought of as an ongoing communication" process. In this context, that ongoing communication was arguably directed at Islamic State leadership and other Indonesian personalities, as much as it may have been directed at Western counterterrorism watchers.

3. Technical Behavioural Content Analysis of Naim's *Teknologi* Chapters

Technical assessments of known and unknown cyber offenders generally start with some characterization of an actor's technical knowledge. We are talking about his or her *native* knowledge, which is experiential and learned, as well as his or her *researched* knowledge, which is knowledge that would be searched for immediately on Google, or in some other manner because the offender is unfamiliar with that tool, for example.

While we cannot know if Naim had native knowledge then or at some point in the past of some of the techniques and programs he wrote about in his e-book, we can question why he included chapters that suggested he demonstrated these step-by-step techniques, when the screen shots found online in this paper indicated he repeatedly did not demonstrate those steps.

Generally, Naim's chapters on hacking read much like an introductory outline of different anonymizing techniques and some of his general opinions on hacking, for example. Naim's chapters throughout a section called *Teknologi* are increasingly filled with screen shots taken from companies and other users online who created those visuals with programs. Because readers see these screen shots accompanied with written instructions, most readers would probably believe that Naim is likely performing these steps on these programs and taking shots. Because of the length of the section *Teknologi*, we will be focusing on content examples that help frame Naim's suggested level or degree of *native* and *researched* knowledge.

3.1 Naim's Plagiarized Screen Shot Tutorials

The examples included below are questionable because there was likely no need for Naim to plagiarize these screen shots, as most of the examples are rudimentary demonstrations; Naim could have simply replicated the same tutorials he appeared to have found online. Further, the examples that appeared to have been plagiarized even include instructions such as entering a username and password, which are arguably laughable when considering that the presumed targeted reader for Naim among the broader Indonesian carding and cybercrime community would have been familiar already with many of these platforms and tools. If we argued that Naim was using other existing content online for his screen shots for operational security, so that he does not include any identifying content, we would have to argue that what he includes in some chapters for the Android may suggest that he is more comfortable using an Android smartphone, which starts to shape his technical profile.

3.2 VPN/OPENVPN

In this early chapter, Naim began including screen shots from what appeared to be other users or content from companies online, rather than demonstrating those steps himself. Naim wrote about calibrating a proxy service like OpenVPN to anonymize yourself online and how important safety is. What appeared to be a plagiarized screen shot demonstrates how a user would set up the OpenVPN installation client and then connect (see **Figure 1** and **Figure 2**).

Naim included the screen shot in Figure 1 in his e-book, which resembles the same screen shot in Figure 2 found by the author in a Google search. The numerous screen shots of OpenVPN and vpnbook as a free proxy vpn service are just informational, helping readers pick a free and secure option among several choices.



Figure 1: Naim's included screenshot of completing an OpenVPN profile installation

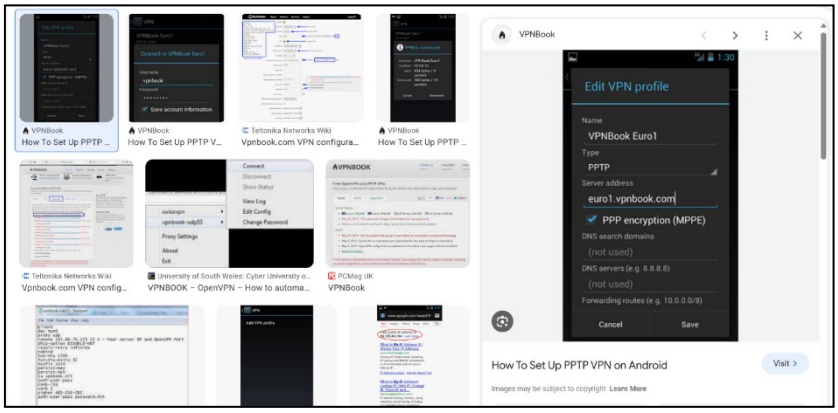


Figure 2: Author's Google search return including the same screenshot

3.3 SSH/OPENSSH

In this chapter, Naim discussed using a Secure Shell Protocol (SSH) cryptographic tunneling program over an unsecure network. Again, Naim used screen shots from existing content online from commercial sites, including screen shots that share some of a desktop Windows background that is likely not his. The same screen shot was found online among initial search results (see **Figure 3** and **Figure 4**).

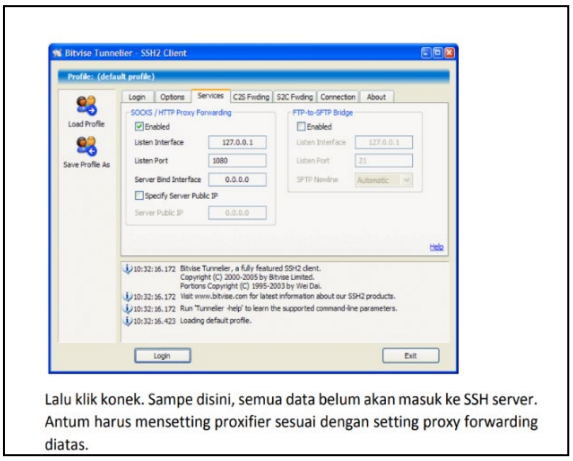


Figure 3: Naim's included screenshot of the connecting with an SSH2 client

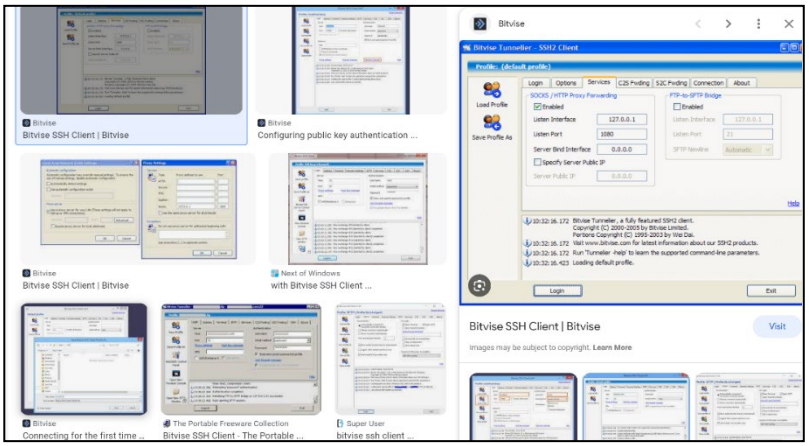


Figure 4: Author's Google search return including the same screenshot

Naim did include additional screen shots in this chapter that demonstrated setting up an SSH tunnel application on an Android, where even in Naim's written content below the screen shots he included port forwarding and setting details for calibrating that application. Again, Naim uses a screen shot from existing content online, in fact, content posted online by another Indonesian-language user on a blog site. The same graphic including the same screen shot that included an Indonesian-language advertisement in the screen shot was found on an Indonesian-language blog detailing the same technique for setting up an Android IP proxy (see **Figure 5** and **Figure 6**).

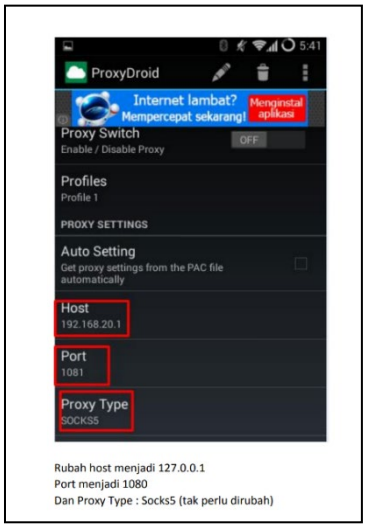


Figure 5: Naim's included screenshot setting up a proxy IP on an Android smartphone

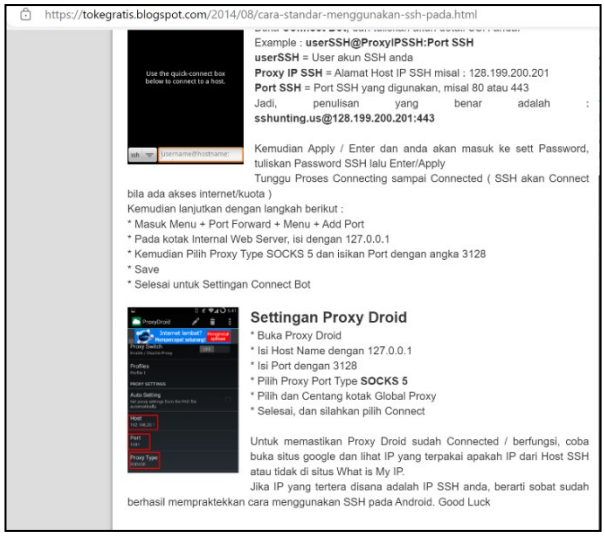


Figure 6: Author's Google search return including the same screenshot from an Indonesian-language blog

3.4 SQL Injection

In this chapter, Naim included an unreferenced screenshot from a blog article published in approximately 2012 (see **Figure 7** and **Figure 8**). While we do not know how Naim searched for this content, the same content is largely still available online when searching for the program names in screen shots that Naim included. There's limited uniformity to the screen shots included throughout these chapters, which suggests that Naim continued to plagiarize these screen shots from other unknown users online. In some screen shots, there are English-language text in different color fonts added to the image. Some of these screen shots included data that could be identifying that appeared simply written over with a marker rather than a highlight to redact that line of text.



Figure 7: Naim's included screenshot demonstrating use of the SQL Poizon tool

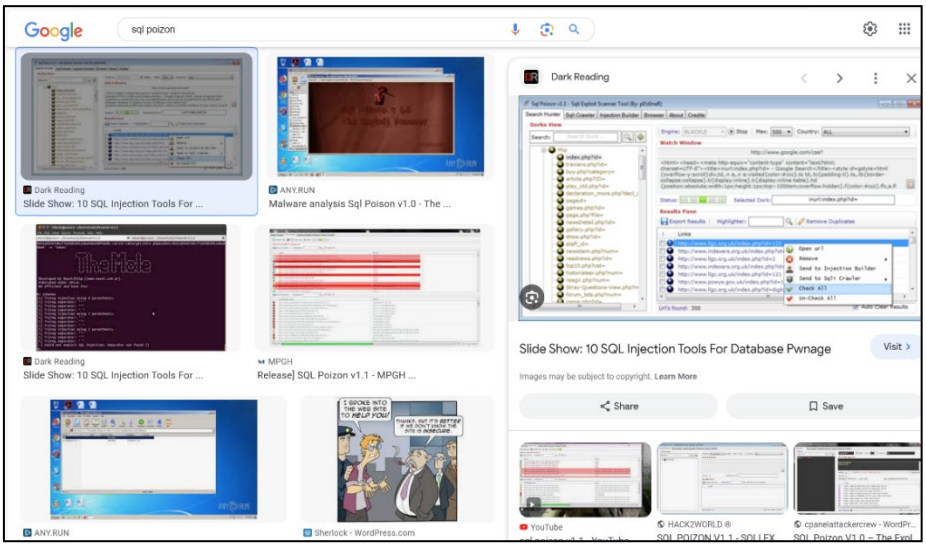


Figure 8: Author's Google search return including the same screenshot

3.5 Naim's Possible *Native* Technical Knowledge

We evaluate an offender's degree of native technical knowledge by examining what we know about that offender's study and other experiences with tools, for example. When there is limited information about what an offender has studied or practiced, we then consider how that offender discusses tools or techniques and what he or she chooses to demonstrate or operationalize in some kind of attack setting.

3.6 Proxy

Naim in this chapter explained what a proxy is and how there are weaknesses to using a free proxy service, rather than paying for a proxy that is not shared and is more secure. When Naim did appear to demonstrate the steps outlined, the process generally involved visiting the website on a desktop or mobile device that provided details on your IP address.

3.7 Socks/VIP72

In this chapter, Naim wrote extensively about both applications, including nuances, such as if you use Microsoft Proxy Server, then you do not need to install WinSock Proxy on the computer you are using, for example. Naim explained how the SOCKS proxy server can relay data safely between the sender and receiver and the connecting server. He noted that SOCKS version 5 added an additional step where each stage of that relay requires authentication. When we see this kind of explanation, there are suggestions that Naim is simply more comfortable writing about this topic because he is experienced using SOCKS, for example.

3.8 Cara Hacking Dan Carding

In this chapter, Naim wrote that he wanted to focus on some of the more common techniques for hacking and carding. First, Naim discussed Structured Query Language (SQL) injection. Naim characterized a SQL injection as a method to "open" a database to collect information in the database as needed. Naim suggested various software programs that could be used to support that technique. Second, Naim talked about exploits, such as Remote Access Trojans (RAT) on mobile phones, that enable collection on the host of that device.

Third, Naim described sniffing or packet sniffing as using a virus keylogger for example to catch packets "on the fly" that could include communications or other important data.[1] Fourth, Naim wrote about cache injections, which is a method he wrote of manufacturing hidden caches of data in files, which he described as a "very easy" technique. Fifth, Naim wrote about Domain Name Server (DNS) poisoning, which he described as a "simple" technique for reconfiguring the origin host DNS to manage control over traffic to the redirected DNS. Naim explained all these techniques in understandable contexts for what appear to be introductory technical readers who may have been unfamiliar with these techniques previously.

[1]Ibid., 90-91

3.9 Web Scam/Scampage

Naim wrote extensively in this chapter about the various *webscam* techniques, including how to manipulate Facebook account domains to trick people. While Naim did include a screen shot of an example of a falsified PayPal URL, this chapter includes lengthy background and plain language examples of how to set up *scampages*.

3.10 Metode Carding

Naim's subsequent chapter on carding methods is even lengthier, organized by platforms where the reader could card victims, such as PayPal and Western Union. Naim included written details in this chapter that describe options for what to input into various platforms to complete illicit carding transactions successfully. The notes Naim included for several listed sites strongly suggested he has some experience using these sites for carding. For example, on a skateboarding clothing site www.draven.com, Naim wrote that the reader can use the shipping address for the billing address and that the reader does not need to use Socks/VPN for this site (see **Figure 10**). There are ten pages of these sites with notes.

3.11 Metode DDOS

Naim in this chapter provided generally available information about Distributed Denial of Service (DDoS) attacks. Naim did include a screen shot of his famed botnet @bahrunnaim_bot and profile photograph as a method of using a botnet to flood a network, however, just above this part of the chapter where Naim wrote about command prompt as another method in Windows, Naim again plagiarized a screen shot demonstrating this method (see **Figure 11** and **Figure 12**).

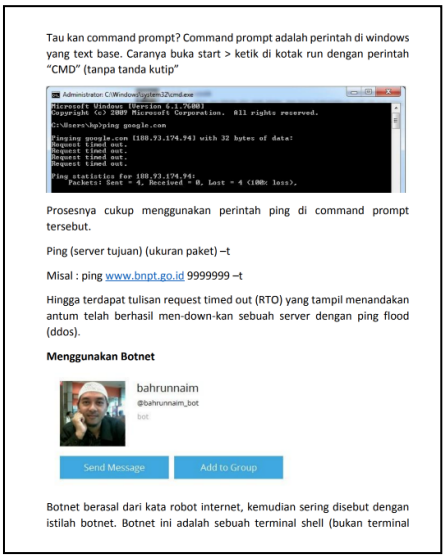


Figure 11: Naim's included screen shot for a command prompt function in Windows

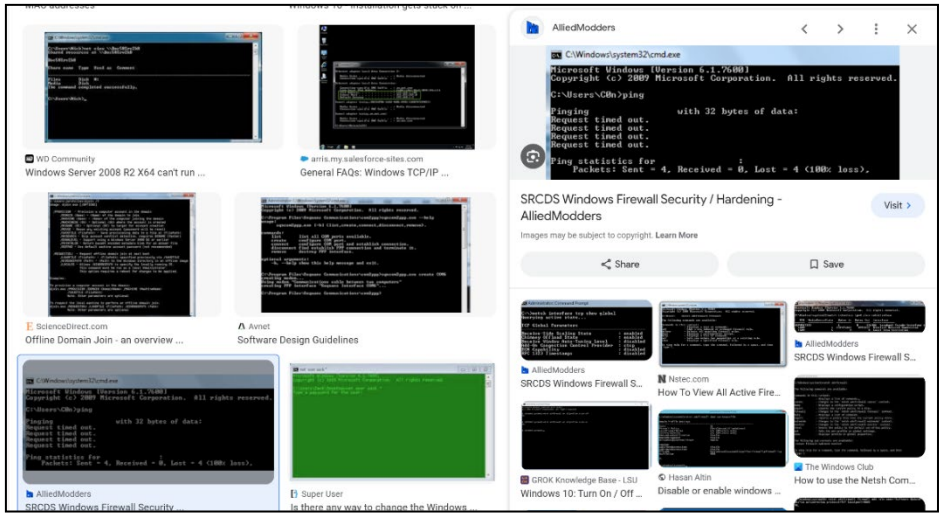


Figure 12: Author's Google search return including the same screenshot

3.12 Naim's Possible *Researched* Technical Knowledge

We evaluate an offender's degree of researched technical knowledge by examining what that offender claims to be able to do related to observable or known experience or training. The preference is to see an offender's behavior regarding a tool or technique, so we can observe either how they demonstrate a tool or technique or what kind of research is required for them to educate themselves on how to use that tool or technique.

For example, if we are observing an administrator who has responsibility for using Powershell in his or her position, but this administrator appears to struggle to appropriately use Powershell and often asks others for help or searches elsewhere for information on how to use Powershell correctly, then that administrator's knowledge of Powershell is likely researched more than native, and we now have some grounding for their technical aptitude or profile. The examples below just based on observation may suggest that Naim has limited researched knowledge, but what does appear to be researched are rudimentary topics that also include some examples of plagiarized screen shots.

3.13 Persiapan: Anonymous

Naim explained early in this chapter that he is trying to educate readers on how to avoid being "tapped" or "bugged" by others online, calling this chapter "preparations". Much of this chapter included Naim writing much like a summary on the hardware used to locate or identify people, such as Global Positioning System (GPS) or the IMEI on a phone. While we are speculating, Naim was likely killed because he did not anonymize appropriately.

3.14 *VPS*

Naim wrote in this chapter about Virtual Private Servers (VPS), such as the benefits of a continuous proxy infrastructure. Naim included a screen shot from a CentOS Webmin version. While the screen shot Naim included was not found during an open-source search, arguably Naim would not want to include the details on his VPS in a screen shot. The open-source search on Google for a similar display from a CentOS Webmin version included many similar screen shots like the one Naim included.

3.15 Google Dork

This 15-page chapter is the longest in Naim's e-book sections on hacking and carding. Most of the pages included lists of search terms to "Google Dork" search for files if the reader has established access on a site or a server to search for files.

3.16 Naim's Only Shared Example of Conducting an Attack Himself

We have found that generally cyber personalities online will showcase or certainly claim some of their exploits or attacks. Part of that learning and teaching process within some of these communities is demonstrating how that offender accomplished a technique and then providing some version of that tool to others so they can do the same. These are reputational displays or signals; however, this is the only example Naim includes in his entire e-book and dozens of chapters about hacking.

3.17 Exploit

In this chapter, Naim discussed the Remote Administrator Trojan (RAT) virus and how to launch this virus from various programs. Naim described how one method of executing this virus is through a website, where the victim sees what appears to be a warning message when they visit the website with a button to return to "safety" to avoid malware on that website. When the victim clicks on the button to return to safety, the virus starts downloading. Naim included a screen shot of an example, which identified www.hendropriyono.com as a malicious site. Naim did not include any further commentary on this website example, but that name is commonly associated with the name of an Indonesian general accused of abuses in East Timor.

In this same chapter, Naim referred to himself as the writer "*penulis*" conducting one of these attacks, for the first time in his e-book. Naim wrote that he remembered one time where an Indonesian counterterrorism officer from a cybercrime division named Roberto had difficulty printing an investigation document. Naim wrote that he "pierced" Roberto's account and observed Roberto's issues with the printer. Roberto's operating system ultimately crashed. Naim wrote that he believed the counterterrorism agency likely kept a backup of system files for Roberto's operating system and other accounts. Naim wrote that he forgot to write a script that would have automatically wiped all the files on Roberto's account.

4. Discussion

This practitioner's paper applied a content analysis of Naim's manifesto chapters on hacking in the context of *native* and *researched* technical analytical frameworks. These are practiced technical analytical frameworks of cyber behavioral analysis that have been applied in many instances, usually involving limited information on known and unknown offenders. We can start to baseline Naim's experiential technical knowledge and capabilities not only based on the content he chose to write about, but demonstrably how he wrote about that content, in contrast to areas he did not write as much about or did not seem comfortable writing about. Additionally, we can gain some sense of Naim's abilities based on the applications and software he chose to write about, whether he demonstrated or plagiarized those tutorial screen shots.

Based just on publicly available information, we have no known record of any instance of Naim conducting computer network operations against any individual or organizational network, other than the story in his manifesto of gaining unauthorized access to an Indonesian counterterrorism official's laptop account. By his own account, he had limited privileges and questionable scripting. Naim appeared to plagiarize most of the tutorial screen shots he included in his manifesto. Randomly sampling one of the tutorial screen shots or diagrams Naim included in another section of the manifesto on bomb making that appeared to include Naim's notes, also appeared to have been plagiarized from another user's content found online during a Google search.

This paper cannot answer why Naim chose to create an e-book rather than communicate this same information the way he had been sharing information on encrypted platforms, or rather than creating video content like many of his peers in the Islamic State. By the time of his death in June 2018, YouTube was already brimming with online tutorials for every kind of hacking technique and platform imaginable; in fact, Naim appeared to plagiarize many of his screen shots from content hosted on Google platforms. Naim shared information about bomb making and weapons on his encrypted platforms, but routinely had to create new accounts in response to his accounts being suspended or banned; perhaps there was greater redundancy in an electronic book that could be more easily shared among broad and niche Indonesian communities and audiences of interest for Naim.

While Naim referred to himself in his manifesto as "just a writer", perhaps he imagined himself to be an Islamic scholar of some note, or he wanted some pedagogical role. This paper suggests it could be all those reasons and more, but arguably Naim's technical pedigree was widely embellished or misunderstood by counterterrorism and cyber analysts. This paper suggests there is greater opportunity to diffuse cyberterrorist influencers by materially highlighting their surprising lack of technical experience and knowledge, rather than simply categorizing them as another concerning terrorist personality. There is a growing world market of cyber talent who even with extremist interests would likely find Naim's plagiarized tutorials laughable.

References

Arianti, Vidia, "Aman Abdurrahman: Ideologue and 'Commander' of IS Supporters in Indonesia," *Counter Terrorist Trends and Analyses* 9, no. 2 (2017).

Arianti, Vidia, "Cybercrime: Financing Terrorism in Indonesia," *RSIS Commentary* 159 (2019).

Baenzer, Marie, Hotspot analysis: use of cyber tools in regional tensions in Southeast Asia, Center for Security Studies, Zurich (August 2018), Cyber-Reports-2018-05.pdf (ethz.ch).

Bliege, Bird, Rebecca, and EricAlden Smith. "Signaling theory, strategic interaction, and symbolic capital." *Current anthropology* 46, no. 2 (2005).

DeAndrea, David C. "Advancing warranting theory." *Communication Theory* 24, no. 2 (2014).

Donath, Judith. "Signals in social supernets." *Journal of computer-mediated communication* 13, no. 1 (2007): 232-233; "Signaling Identity." *Sociable Media Group* 10 (2007).

Gunaratna, Rohan, "Mastermind of Terror: The Life and Death of Bahrun Naim," *Counter Terrorist Trends and Analyses* 10, no. 10 (2018).

International Institute of Counter-Terrorism, "IS-Supporting Hacktivists in Southeast Asia", Reichman University, June 2, 2019, 10, IS-supporting Hacktivists in SE Asia v1 (ict.org.il).

Joscelyn, Thomas, "Indonesian authorities hunt Islamic State operative's cyber recruits," *FDD's Long War Journal*, April 18, 2017, www.longwarjournal.org/archives/2017/04/Indonesian-authorities-hunt-islamic-state-operatives-cyber-recruits.php.

Kassim, Y. Z., "The Jakarta assault: pre-empting the rise of IS Indonesia," *RSIS Com mentary* 17 (2016).

Lang, Martin, Radim Chvaja, and Benjamin Grant Purzycki, "Costly commitment signals promote co-operation and sacrifice during intergroup conflict", (2023).

Liu, Christine M., and Donath, Judith. "Urbanhermes: social signaling with electronic fashion." In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 1. 2006.

Mulia, Musdah, "Bahaya radikalisme dan kekerasan ekstrismisme," *AL-WARDAH: Jurnal Kajian Perempuan, Gender dan Agama* 12, no. 2 (2019).

Naim, Bahrun. *"Sebuah Perjalanan, Rahasia".*
Romadlan, Said, "The Discourse of Meaning of Jihad in Muhammadiyah Circle (A Hermeneutics Perspective)," *Komunikator* 11, no. 2 (2019).
Rozika, Weldi, "Propaganda dan penyebaran ideologi terorisme melalui media internet (Studi kasus pelaku cyber terorisme oleh bahrun naim)," *Jurnal Ilmu Kepolisian* 11, no. 2 (2017).
Schulze, Kirsten E., "The Jakarta attack and the Islamic State threat to Indonesia," *CTC Sentinel* 9 (2016).
Singh, Bilveer, "Southeast Asian jihadi leaders in the post-Marawi era," *RSIS Commentaries, CO18007* 16 (2018).
Taufiqurrohman, Muh, and Ardi Putra Prasetya, "A Rising Indonesian Jihadist Plotter: Bahrun Naim," *Counter Terrorist Trends and Analyses* 8, no. 11 (2016).

Typology of State Actors' Behavior in Cyber Space

Ada Peter and Ujunwa Ohakpougwu
Communications Department, Covenant University, Nigeria

ada.peter@covenantuniversity.edu.ng;
ado909@g.harvard.edu
ujunwa.ohakpougwupgs@stu.cu.edu.ng

Abstract: Cyberwar is no longer subject to "if" but "when." Despite growing interest in cyberspace and cyber war readiness and resilience among academics, researchers, policymakers, and the media, the area needs to be more robust with different terminology that strategically captures the activities of state actors in cyberspace. The paper aims to provide a strategic classification of the activities of state actors in cyberspace. A typology is developed to encapsulate the strategic complexity of the activities of state actors in this terrain. The typology can illuminate the current global proportion and payoff of each type.

Keywords: Cyber, Cyber war, State actors, Behavior, Strategy, Cyberspace, Replicator dynamics

1. Background

We have intercepted a code that looks like it may be designed to be triggered in the next few weeks – basically during the swearing-in period for the United States Congress and state governments across the country, including the white house and Capitol Hill, and precisely January 3 when Vice President Kamala Harris will be administering the oath of office. Wondering why? Russia is seeking a sizable retaliation point for massive U.S. support to Ukraine which led to its recent withdrawal from Kherson. Russians think the best way to send this message and disrupt the government at critical moments in Washington and across the states is to turn out the lights and hamper the transition period requiring the focus of Washington, D.H.S., and maybe D.O.D. Who knows what they will try to pull on U.S. allies?

The situation above is entirely fictional, but the events are drawn from headlines and current affairs to reflect that in cyberspace, the probability of warfare through the fifth domain has shifted from questions of "if cyberwarfare occurs" to questions of "when cyberwarfare occurs." Hence there are research attempts to examine how cyberspace evolved from an intelligence tool to a destructive weapon targeted at critical infrastructures (Peter & Ohakpougwu, 2023). There are also several efforts to measure how well state actors are preparing, prepared, or at risk (Jegede, 2019 et.al; Peter & Sobowale, 2015.) These existing assessments estimate and evaluate the cyber power and, by extension, the cyberwar readiness of state actors.

Among several assessments, the World Economic Forum's 2016 *guesstimation* suggests that the United States, China, Russia, Israel, and the United Kingdom had the most developed cyber warfare capabilities. The International Telecommunication Union (I.T.U.) Global Cybersecurity Index (GCI), which assesses the commitment of member states to cyber security along five pillars, rates the United States, United Kingdom, France, Russia, and China's performance between 80% -100%, with the U.S., rated 100% (I.T.U., 2014,2017,2018, 2020). The Cyber Resilience Preparedness Index (C.R.I.), published in 2017, zoomed in on Africa's top 12 emerging economies. Egypt, Nigeria, and Kenya were classified as the most cyber-resilient countries on the continent (Peter, 2017). The Belfer Center's National Cyber Power Index (NCPI) at the Harvard Kennedy School considered cyber power in the context of seven national objectives countries pursue using cyber means (Voo *et al.,* 2020). Again, the United States, the United Kingdom, China, the Netherlands, Russia, and France were ranked the top six countries with the highest cyber power. Another measure worthy of note is the International Institute for strategic studies (IISS) qualitative assessment of 15 countries' cyber power. The only country classified as a tier 1 country on the IISS assessment is the United States. The United Kingdom, France, Russia, China, Iran, and DPRK were classified as tier 2 countries (IISS 2021).

At first glance, the position of high-ranking countries on each assessment subtly projects a false sense of cyberwar preparedness, readiness, and resilience to the intent and capabilities of cyber adversaries. However, state actors resting or acting on their GCI, IISS, C.R.I., and NCPI performance can be likened to ticking off as available all the weapons necessary in the outbreak of war without reflecting on how well weapons will or should perform in the face of adversity or a more sophisticated weapon. Also, the various cyber power indices must provide high-level defense decision-makers with appropriate information for assessing cyberwar readiness. For instance, countries like the United States, United Kingdom, France, Russia, and Germany, which rank 98%-100% on the GCI, can neither rest nor double their efforts based on their GCI ranking. The reason is that neither

nation's high or low commitment reflects cyber adversaries' interest or capacity to launch cyberattacks on a country's cyber infrastructure. Also, though the United States, United Kingdom, France, Russia, and Germany rank 98%-100% on the GCI, critical infrastructures belonging to these top-ranked countries remain at risk. Compromise, espionage, malware, ransomware, and attacks on industrial control systems continue to increase because cyber warfare tools, tactics, and procedures (TTP) evolve so quickly that cyber power and capability measures based on cyber defense legislation and law enforcement remain behind the activity curves of state and non-state attackers.

Hence, a replicator dynamic behavior model for assessing a nation's cyberwar readiness and resilience may be a more helpful framework for state actors to track each actor's cyber resilience, cyberwar readiness, and possible future behavior. However, to develop a cyber replicator dynamic equation of state behaviors in cyberspace, the research requires three critical pieces of data. First is some type of behavior in cyberspace, next is the proportion of each type, then the payoff for each type of cyber behavior within and between state actors.

The current paper, therefore, focuses on classifying the varying means or processes by which nations prepare and pursue cyber warfare across 16 critical infrastructures dependent on cyberspace. Typologies are critical for three reasons. The first is to encapsulate the strategic complexity of the activities and behavior of state actors in cyber terrain. Second, it serves as a precursor for illuminating each type's current global proportion and payoff, which are necessary to examine future evolutionary dynamics. Moreover, as a valuable tool for indexing, a typology of state actors' behavior in cyberspace holds insights into each type of state activity's evolutionary possibilities. It provides opportunities for close strategic case studies of individual state actors' behavior in cyberspace.

Subsequent papers report the procedures for identifying the proportions of each type, the payoffs, and an index of each nation's performance based on each type. With data on these critical aspects, the feasibility of a dynamic replicator model of state actors' behavior in cyberspace increases so that decision-makers can step away from dominantly reactive cyber strategies to proactive approaches.

2. Method

To create the typology of state actors' behavior in cyberspace, the research adopts Yin's cross-case/multiple-case studies approach to examine multiple cases of state-backed cyber-attacks. The design for developing a typology of state actors' practices in cyberspace is based on a replication logic. According to Yin, reliance on theoretical concepts remains one of the essential strategies for completing successful case studies, whether exploratory, descriptive, or explanatory (Yin, 2017). These include the development of preliminary concepts at the outset of a case study, placing it within the appropriate research literature defining the units of analysis, identifying criteria for selecting and screening potential cases, and suggesting variables of interest. Donella Meadows's systems theory also provides the theoretical framework guiding the questions used to cross-examine each case (Meadows, 2008). The theory suggests that the behavior of state actors in cyberspace can only be understood by studying trends over time rather than focusing attention on individual events. At a 90% confidence level, the researcher selected 32 out of 776 state-backed attackers executed by the top five state actors responsible for 84% of all the attacks between 2005 and 2021 (See Figure 1). The top five state actors responsible for 84% of all attacks include the United States, Russia, China, North Korea, and Iran. The criteria for selecting the cases are

- One of these five countries (United States, Russia, China, North Korea, and Iran) executed the attack on another state actor between 2005 and 2021
- The attacks with the longest recovery time executed by these countries
- The attacks must be on cyber-dependent critical infrastructures that support the functions of other critical infrastructures.

The study excludes the behavior of state actors in contexts of authorized access for intelligence gathering and sharing like those that exist among the five or seven eyes.

It is also important to note that 85% of these attacks occurred between countries experiencing tense or sour political, economic, and social relations at the time of the attack. Most US-reported cyber-attacks were towards Russia, Iran, and North Korea.

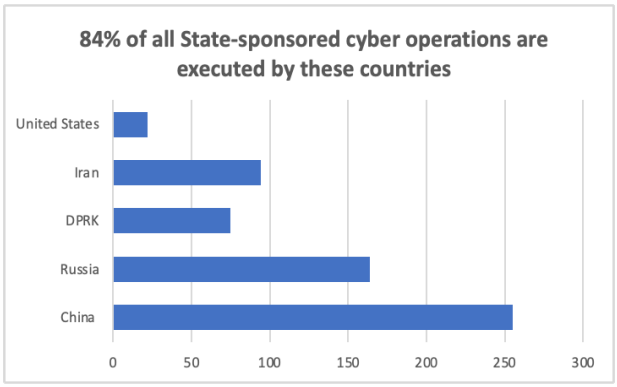


Figure 1: Five State Actors responsible for 85% of State-backed Disruptions in Cyberspace

Based on these criteria, 5% of the attacks with the most significant impact on critical infrastructure and severity executed by these countries were selected. Impact and severity were measured by how many critical sectors depended on the affected sector, the recovery time, and the financial loss. Analysis of 16 cyber-dependent critical sectors by Peter (2022) reveals that some sectors are more dependent on other sectors (See figure 2), so a successful attack on a sector with many dependencies yields greater rewards for the attacker. The longer the time it takes to recover, the higher the severity. The research objects for the study were the targeted cyber-dependent critical infrastructures associated with it political, social, and historical relevance and or issues.

Country	Target Rate	Cases studied
China	255	13
Russia	164	8
DPRK	75	4
Iran	94	5
United States	22	2

We then conducted a traditional cross-case analysis of 32 cases of state-backed attacks collected from the Council on Foreign Relations data on state-sponsored cyber operations (https://www.cfr.org/cyber-operations). Famous examples of the 32 cases selected include Stuxnet; the Microsoft Exchange server attack, which affected the European Banking Authority and other organizations globally; the SONY hack; the U.S. Colonial Pipelines, the Russian electric grid; the SolarWinds attack, etc**.** The following key questions were used to examine each case.

- What critical infrastructure was attacked? EXECUTE
- Does the attack share common conditions with other attacks?
- What are the shared common and uncommon conditions?
- Which communication or information technology aided the attack? (Satellite, Network services, Network hardware products, cloud services, cybersecurity tools, social networks, internet) FARMING
- At the time of the attack, what was the state of the communication/I.T. infrastructure that aided the attack? A home or foreign-owned Communication and I.T. infrastructure? (FARMING)
- How did the attack work? (EXPLOIT)
- What actions did the various actors take (hero and villain) before and after the attack? -DEV/PATCH
- What is the geographical reach of the communication/I.T. infrastructure that aided the attack?
- Has the communication/I.T. infrastructure that aided the attack been updated in the past?
- What are the diplomatic relations between the hero and the villain? (Economic, social, intelligence sharing, and political)

The researcher also reviewed several contents and synthesized empirical findings, public debates, and reports mentioning state actors. These include N.S.A. leaked documents, proceedings of vulnerability conferences

(DEFCON, BruCON, Black Hat Europe, Cold Blue, Null CON, Zero Nights, 44CON), vulnerability reports, exploit reports, patches, and the CISA database. The review focused only on state actors' offensive and defensive activities across 16 cyber-dependent critical infrastructure sectors (See figure 2), including Chemical, Commercial, Communication, Defense Industrial Base, Dams, Energy, Emergency Services, Financial, Food and Agriculture, Government Facilities, Health and Public Health, Information Technology, Critical Manufacturing, Nuclear, Transportation, Water, and Wastewater Sectors.

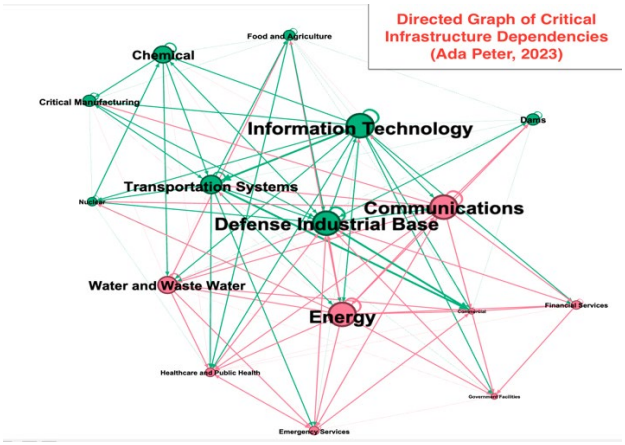


Figure 2: Directed Graph of Critical Infrastructure Dependencies (Ada Peter, 2023)

The figure represents 16 critical infrastructure sectors (nodes) of cyber-dependent States. Arrows pointing to a node show the sectors depending on it and vice versa. The larger nodes (e.g., defense industrial sector) indicate that more sectors depend on that node than the smaller nodes (e.g., chemical sector). Thus, a targeted disruption or destruction of the larger nodes will have ripple effects on all the dependent sectors, generating a higher impact than an attack on the smaller nodes. The larger the node, the higher the risks; thus, decentralizing critical infrastructure dependency within and between sectors may enhance cyber resilience outcomes.

The preceding data analysis broadly revealed six means or processes by which nations prepare and pursue cyber warfare across 16 critical infrastructure sectors (See Figure 2) dependent on six critical areas of communication and information technology. The means and processes identified and classified include

- Farming vulnerabilities
- Unearthing vulnerabilities
- Developing payload exploits
- Executing exploits
- Storing arsenals of allies' and enemies' vulnerabilities and possible exploits
- Patching vulnerabilities

Farming Vulnerability (F.V.): State actors understudy were found to use the spread of homemade and foreign technologies as a cyberweapon. F.V. appears both as an offensive and defensive posture. F.V., on the one hand, gives the home country offensive access to foreign data and, on the other hand, offers offshore countries the opportunity to gain access to the home country's network services, equipment, intellectual property, how it works, understand the vulnerabilities, and build ready-to-deploy exploits. F.V. is the necessary means by which state actors pursue cyber war. Peter & Ohakpougwu's (2023) assessment of how the internet gets weaponized, indicates that farming vulnerability occurs across all the stages of weaponization, excluding cyber-independent phase of cyber weaponization.

Farming refers to the spread of state actors' information and communication technology services and products, including cloud security, network services, and equipment, satellite services and equipment, and internet services and equipment to other countries and continents. Examples include backbone I.S.P.s such as Backbone I.S.P.s (AT&T, CenturyLink, Cogent Communications, Deutsche Telekom, Global Telecom and Technology (GTT), N.T.T. Communications, Sprint, Tata Communications, Telecom Italia Sparkle, Telia Carrier, and Verizon), Cloud service (Amazon Web Services (A.W.S.), Microsoft Azure, Google Cloud Platform (GCP), Alibaba Cloud, Oracle Cloud, I.B.M. Cloud (Kyndryl), Tencent Cloud, OVHcloud, DigitalOcean, and Linode (owned by Akamai)—hardware and software companies like Apple, Google, Weibo, WeChat, TikTok, Huawei, etc.

Unearthing Vulnerabilities (U.V.): The country's understudy has commissioned teams whose singular goal is to unearth internal and external vulnerabilities. U.V. is both an offensive and defensive posture. U.V. refers to countries coordinating the daily activity of finding their vulnerabilities and external vulnerabilities of on-shore and offshore software and hardware network services and equipment, cloud security, internet services, and satellites used across its critical infrastructure sectors, including Chemical, Communication, Commercial Defense industrial base, Energy, Finance, Government facilities, Health and public health services, Information technology, Manufacturing, Nuclear, Transportation, Water, and wastewater. For instance, the primary purpose of the United States' Common Vulnerabilities and Exposures (C.V.E.) The program identifies vulnerabilities uniquely and associates specific versions of code bases (e.g., software and shared libraries) to those vulnerabilities. China's version of the US NVD is the **Chinese National Vulnerability Database** (**CNNVD**). It is the national vulnerability database that catalogs all software defects and patches. In the cyber weaponization process, countries are only capable of unearthing vulnerability particularly in building resilient critical infrastructure phase (Peter & Ohakpougwu, 2023).

Developing Exploits (DEVEX): State actors also develop exploits. For adversarial state actors, finding or unearthing vulnerabilities is worthless without payloads to exploit those vulnerabilities. While finding vulnerabilities are worthful for adversaries with payloads to exploit those vulnerabilities, on the flip side, exploits are worthless for the adversaries when those vulnerabilities are quickly patched. Hence, state actors step beyond identifying common application vulnerabilities to developing exploits that penetration testing may not capture. Publicly available examples of state-backed efforts to develop exploits are ethical hackers, more intrusive and involve active exploitation of security vulnerabilities. The Cybersecurity and Infrastructure Security Agency (CISA), a federal agency of the U.S. government, recently selected Bug Crowd to launch its first federal civilian enterprise-wide crowdsourced vulnerability disclosure policy (V.D.P.) platform. The V.D.P. platform enables agencies to identify and monitor vulnerabilities in critical systems by receiving security feedback from uniquely skilled ethical hackers worldwide. Two critical cycles must be completed for these ethical hackers to provide feedback. First, identify the vulnerabilities, and second, exploit the vulnerabilities. N.S.A. leaked documents also reveal how the United States exploits vulnerabilities for military, economic, or political motives. More so, intelligence gathering and espionage via digital means are impossible without state actors using developed exploits to gain access back door access to information. While several nations are engaging DevEx, as Table One indicates, China develops payloads for exploiting known and unknown vulnerabilities.

Executing Exploits (EXEX): The countries' understudy has executed detrimental exploits at different times. These are the most common, repeated, and reported means of pursuing cyber warfare. As an offensive posture, executing exploits include state-backed attacks on the digital infrastructure of other countries. Several state actors in North America, Asia and Europe have shown traits of these behavior. Examples include the 2022 disruption of U.S. gas pipelines, Russia's alleged use of social media campaign to interfere in the 2016 U.S. presidential elections, the late 2014 North Korean attack on Sony Pictures in connection to a planned release of the poorly reviewed movie, *the interview,* the 2010 American Stuxnet attack on Iran and North Koreas weapon's program and the Chinese decades-long espionage of U.S. trade secrets including those extracted *from the* Booz Allan Hamilton, Russian disruption of 2016 U.S. elections, and the Chinese-backed attack on the European banking authority. These attacks are simply apparent execution of identified vulnerabilities and developed exploits. The benefits and gains can range from the show of cyber power to political, military, economic, and diplomatic benefits. The Centre for Foreign Relations tracks reports of this type of state-backed attacks. Unlike other continents, state actors in Africa, have no records of executing disruptive exploits (Peter & Sobowale, 2015). Non execution of exploits also suggests that state actors in Africa may be lagging around the pre-cyber dependent critical infrastructure phase (Peter & Ohakpougwu, 2023).

Patching (P): State actors commission patches of vulnerable cyber tools. Patching is a defensive posture. It is essential to coordinate vulnerability disclosure, especially for public-facing digital infrastructure. It refers to fixing vulnerabilities identified within a digital infrastructure as soon as possible. State actors with patching standards and laws are better defenders of their digital infrastructure since patches incapacitate payloads configured to exploit known-known and known-unknown vulnerabilities.

Storing Vulnerabilities and Exploits (SVEX): Countries store vulnerabilities and exploits. Storing vulnerabilities, especially zero-day vulnerabilities, can be likened to storing a war arsenal that will be deployed when needed. However, since storing vulnerabilities and exploits to pursue cyber warfare is highly correlated to farming and unearthing, DEVEX, and EXEX, the model assumes that countries that rank above average for all five certainly store or stockpile zero-day vulnerabilities. Zero-day stockpiles are treasures unearthed and stored for valuable times. Governments' use of zero-day exploits has exploded over the last decade, feeding a lucrative market for

defense contractors and others who uncover critical flaws in the software (and hardware) and sell information about these vulnerabilities to their government customers. For example, the infamous Stuxnet, a digital weapon used to attack Iran's uranium enrichment program, used four zero-day exploits. While many state actors in Asia, Europe, Australia and North America have and continue to exhibit the behavior of storing arsenals, the tendencies of state actors in Africa to do same is not unclear since some state actors on the continent are gradually approaching the full spectrum of a cyber dependence (Peter & Sobowale, 2015).

3. Preliminary Conclusion

The preceding six means or processes by which nations prepare and pursue cyber warfare across 16 critical infrastructures dependent on cyberspace will constitute the basis for developing a dynamic replicator model accessing the evolution and performance of each state across each means. We hypothesize that state actors involved in all six means in an evolving approach are cyber-war-ready, while state actors passively or actively involved in some but not all means are in cyber-war readiness momentum, others without a trait of these activities are unprepared, their network readiness ranking notwithstanding.

Acknowledgement

We recognize the conference participation support from Covenant University

References

44CON https://44con.com Accessed November 16, 2022

Black Hat Europe https://www.blackhat.com/eu-22/ Accessed November 16, 2022

BRUCON https://www.brucon.org/2022/ Accessed November 16, 2022

Council on Foreign Relations Cyber Operations Tracker (2022). https://www.cfr.org/cyber-operations Accessed September 25, 2022

Cybersecurity and Infrastructure Security Agency (CISA) https://www.cisa.gov Accessed October 5, 2022

DEF CON (2022). https://defcon.org Accessed November 16, 2022

Donella Meadows (2008). Thinking in Systems. Chelsea Green Publishing

International Institute for Strategic Studies. Cyber Capabilities and National Power: A Net Assessment (2021). https://www.iiss.org/blogs/research-paper/2021/06/cyber-capabilities-national-power

I.T.U. (2014) https://www.itu.int/pub/D-STR-SECU-2015 Accessed September 5, 2022

I.T.U. (2017) https://www.itu.int/pub/D-STR-GCI.01-2017 Accessed September 5, 2022

I.T.U.(2018)https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-GCI.01-2018-PDF-E.pdf Accessed September 5, 2022

I.T.U. (2020) https://www.itu.int/en/ITU-D/Cybersecurity/Pages/global-cybersecurity-index.aspx Accessed September 5, 2022

Jegede, A., Okorie, N., Oyero, O., & Olowolafe O. (2019). Post humanism, Virtual Warfare, and the Defence Preparedness of Nations: A case for Africa's readiness.
(2019) *African Renaissance*, 16 (3), pp. 67-90.

Julia Voo, Irfan Hemani ,Simon Jones, Winnona DeSombre, Daniel Cassidy, Anina Schwarzenbach. National Cyber Power Index 2020 Methodology and Analytical Considerations (2020). https://www.belfercenter.org/sites/default/files/2020-09/NCPI_2020.pdf Accessed September 5, 2022

Marco De Falco (2012). https://ccdcoe.org/library/publications/?search=Stuxnet October 5, 2022

NIST (2022) National Vulnerability Database. https://nvd.nist.gov Accessed September 5, 2022

Peter, A. (2017). Cyber Resilience Preparedness of Africa's Top Emerging Economies (June 10, 2017). International Journal of Critical Infrastructure Protection, Vol.17 49-59

Peter, A. & Ohakpougwu, U. (2023). Origins of Cyberwarfare: How the Internet got Weaponized. (2023) *Proceedings of the 10th European Conference on Social Media, ECSM 2023*, pp. 364-372

Peter, A., Sobowale, I. (2015). Proceedings of the 25th International Business Information Management Association Conference - Innovation Vision 2020: From Regional Development Sustainability to Global Economic Growth, IBIMA 2015, 2015, pp. 73–86

Robert Yin (2017). Case Study Research and Applications Design and Methods. Sage Publications

Friend or Foe – The Impact of ChatGPT on Capture the Flag Competitions

Heloise Pieterse
Council for Scientific and Industrial Research, Pretoria, South Africa

hpieterse@csir.co.za

Abstract: ChatGPT, an artificial intelligence (AI)-based chatbot, has taken the world by storm since the technology's release to the public in November 2022. The first reactions were awe and amazement as ChatGPT presented the capability to instantly respond to various text-based questions following a conversational approach. However, it is ChatGPT's ability to complete more advanced tasks, such as supplying source code to programming-related questions or generating complete articles focusing on a specific topic, which has caused eyebrows to be raised. The capabilities offered by ChatGPT, fuelled by popularity and easy accessibility, have introduced several new challenges for the academic sector. One such challenge is the concept of AI-assisted cheating, where students utilise chatbots, such as ChatGPT, to answer specific questions or complete assignments. Although various research studies have explored the impact of ChatGPT on university education, few studies have discussed the influence of ChatGPT on Capture the Flag (CTF) competitions. CTF competitions offer a popular platform to promote cybersecurity education, allowing students to gain hands-on experience solving cybersecurity challenges in a fun but controlled environment. The typical style of CTF challenges usually follows a question-answer format, which offers students the ideal opportunity to enlist the assistance of ChatGPT. This paper investigates the ability of ChatGPT to assist and aid students in solving CTF challenges. The exploratory study involves past CTF challenges across various categories and the questioning of ChatGPT in an attempt to solve the challenges. The outcome of the study reveals that although ChatGPT can assist students with challenges during CTF competitions, the assistance that can be offered is minimal. Instead of producing answers to CTF challenges, ChatGPT can merely offer insight or guidance regarding the questions asked.

Keywords: ChatGPT, Artificial Intelligence, Education, Cyber Security, Capture the Flag, CTF

1. Introduction

One of the first thoughts that often comes to mind when considering Artificial Intelligence (AI) is the Terminator, a complex and intelligent machine capable of perceiving and understanding its surroundings to take appropriate actions (Wooldridge and Jennings, 1995). In literature, however, AI is defined as the study of developing programs or computers capable of performing tasks that would require human intelligence (Boden, 1996). The utilisation of AI to enhance software, hardware implementations, and applications in the 21st century has witnessed steady progress and continuous adoption. Various sectors that are benefiting from the advances in AI include healthcare (Davenport and Kalakota, 2019), cyber security (Zhang et al, 2022), as well as education (Zawacki-Richter, 2019). With the November 2022 release of ChatGPT, a conversational chatbot created by OpenAI (OpenAI, 2023), various new opportunities and challenges emerged for the academic sector.

ChatGPT is a variation of the Generative Pre-trained Transformer 3 (GPT-3), also called GPT-3.5, which is a form of generative AI based on the Transformer architecture (Vaswani et al, 2017) and pre-trained using a large corpus of unlabelled text (Radford et al, 2018). The capabilities of GPT-3 appear to be endless and offer numerous opportunities to the academic sector. From a teaching perspective, GPT-3 can assist with the preparation of lessons, translation of educational materials, as well as generate assessments, tasks, and quizzes (Kasneci et al, 2023). GPT-3 also presents opportunities for learning, such as providing feedback to questions in real-time, assisting with the writing of research assignments, generating summaries of long documents, and enabling the development of programming skills (Kasneci et al, 2023; Lo, 2023). It is such capabilities, which have been successfully demonstrated by ChatGPT (Malinka et al, 2023) that have introduced new challenges for the academic sector.

A current challenge for the academic sector is the concept of AI-assisted cheating. Most, if not all universities have policies in place emphasizing that cheating in any shape or form, such as plagiarism or paraphrasing without the proper citation, is not permitted (Anders, 2023). Students have and will always be expected to do their own work. However, AI technologies, such as ChatGPT, are now freely available and accessible to any student with a smartphone and Internet connectivity. While various research studies have explored the impact of ChatGPT on the academic sector (Lo, 2023; Malinka et al, 2023), very few studies have discussed the influence of ChatGPT on Capture the Flag (CTF) competitions.

Since the inception of the DEFCON CTF in 1996 (CTF History, 2023; Raman et al, 2014), CTF competitions have grown exponentially and continue to promote cybersecurity education (von Solms and van Heerden, 2022) through creative, time-sensitive, puzzle-solving challenges that focus on a specific cybersecurity skill (e.g., web exploitation, forensics, cryptography, forensics, etc.) (Carlisle et al, 2020). Although CTF competitions can follow one of three formats, namely jeopardy-style, attack-defence or mixed, most CTF competitions adopt the jeopardy-style format (Tatsuya and Rei, 2023). Jeopardy-style CTF competitions follow a question-answer format, where students are required to solve various cybersecurity challenges and obtain the answer, which is often called the flag (Carlisle et al, 2020; Raman et al, 2014). Such a format offers students the ideal opportunity to enlist the assistance of ChatGPT.

Two research studies have investigated the potential impact of ChatGPT on CTF competitions. Malinka et al. (2023) focused on computer security-oriented specialisation and evaluated the influence of ChatGPT regarding the completion of examinations, programming assignments, term papers, and interactive projects (such as CTF challenges). Concerning the interactive assignments, the study found that ChatGPT is unable to solve such assignments independently but can only guide the students towards online tools that can possibly assist in revealing the answer (Malinka et al, 2023). Satake and Otsuka (2023) experimented with ChatGPT and attempted to solve the CTF challenges of the picoCTF 2022 competition. Using ChatGPT, the authors were able to obtain 48 flags from 64 challenges (Tatsuya and Rei, 2023). These research studies have produced contradictory results and the impact of ChatGPT on CTF competitions remains uncertain.

This paper builds on the previous research studies conducted and investigates the ability of ChatGPT to assist and aid students in solving CTF challenges. An exploratory study is conducted that involves past CTF challenges across various categories and the questioning of ChatGPT in an attempt to solve the challenges. The research question of the exploratory study is: How can ChatGPT help students solve CTF challenges? Therefore, the purpose of this study is to examine the assistance ChatGPT can offer to students participating in CTF competitions and whether structural changes to CTF challenges will be required.

The remainder of this paper is structured as follows. Section 2 outlines the evolution of chatbots, discussing the various GPT versions as well as chatbot alternatives. Presented in Section 3 is the purpose of CTF competitions. Section 4 presents the methodology followed to conduct the exploratory study. In Section 5, various CTF challenges are attempted to be solved using ChatGPT and the results are discussed. Section 6 concludes the paper.

2. Evolution of Chatbots

As early as 1950, Alan Turing, pondered whether a machine could demonstrate intelligent behaviour, indistinguishable from that of a human (Turing, 2009). Today, this concept is known as the Turing test and is considered by many as the foundation for the development of chatbots. The purpose of any chatbot is to simulate and maintain a human conversation using natural language (Shawar and Atwell, 2007). Such ability was first exhibited by ELIZA, a chatbot developed in 1966 that enabled natural language conversation with a computer (Weizenbaum, 1966). While chatbots are no longer a novel concept, such technology has been revolutionised with the development and release of the Generative Pre-trained Transformer or GPT.

2.1 Generative Pre-Trained Transformer

The first version of GPT (GPT-1) was released in 2018 by OpenAI (Radford et al, 2018) and demonstrated the successful construction of a large language model following a semi-supervised approach that used a combination of unsupervised pre-training and supervised fine-tuning (De Angelis et al, 2023). The capabilities of the model included generating new text (e.g., translation or summarisation) and answering questions (Kocoń et al, 2023).

With the release of GPT-2 in February 2019, the number of parameters used to train the model increased from 117 million to 1.5 billion. The increase in parameters improved the text generation capabilities of GPT-2, producing more coherent, multi-paragraph text. Due to the potential misuse of GPT-2, the model was not initially released to the public (Marr, 2023).

Announced in June 2020, GPT-3 presented a significant leap forward for the model, which was trained using an incredible 175 billion parameters. GPT-3 demonstrated advanced text-generation capabilities, such as drafting emails, creating poetry, and even producing programming code (Marr, 2023). While powerful, GPT-3 still exhibited unintended behaviour, including generating biased text, creating fictitious facts, or failing to follow user instructions (Ouyang et al, 2022).

InstructGPT was developed to improve the responses of the GPT-3 model to questions supplied by humans and relied on Reinforcement Learning from Human Feedback (RLHF) to update the model parameters. The InstructGPT models showed improvements regarding the truthfulness and accuracy of text generation, however, simple mistakes were still detected (Ouyang et al, 2022). Regardless, InstructGPT formed the foundation for ChatGPT (Kocoń et al, 2023).

ChatGPT's successor, GPT-4, continues to exponentially improve the model's capabilities, which now support image input, lower probability of generating offensive outputs, increased factual accuracy, as well as the ability to search the Internet in real-time (Kocoń et al, 2023). However, the rise of ChatGPT has stimulated the development of other chatbots.

2.2 Rise of the Chatbots

The improvements of generative AI for chatbot technology have advanced significantly in the past year causing a new AI arms race to arise. While ChatGPT appears to be the current leader, Bard, Bing Chat and Ernie have emerged as strong competitors.

Bard, originally called Meena (Adiwardana et al, 2020), is based on a lightweight and optimised version of LaMDA (Language Model for Dialogue Applications), which closely resembles most Large Language Models (LLMs) used today (Manyika, 2023). The chatbot was pre-trained using various publicly accessible data sources and further improved using RLHF. Since the release of Bard, the chatbot has undergone various refinements and improvements (Adiwardana et al, 2020), however, Bard's responses might still be inaccurate, especially when questioning about complex or factual topics, or reflect bias and subjective responses (Manyika, 2023).

Bing Chat, which forms part of the Bing search engine, is based on GPT-4 and addresses some of the weaknesses associated with ChatGPT. The noteworthy difference between Bing Chat and ChatGPT is Bing Chat's ability to access the Internet and yield more updated answers. While the answers produced by Bing Chat are short and concise, supplied references reveal the answers are collated exclusively from non-academic sources (Rudolph, Tan and Tan, 2023).

The current AI leader among Chinese corporations is Baidu, who first released a GPT-3 equivalent, called the *Enhanced Representation from kNowledge IntEgration* (Ernie) 3.0, in 2019. A few years later, Baidu announced its text-to-image model commonly called Erinie-VILG. However, it was with the release of Ernie Bot in March 2023 that a Chinese rival for ChatGPT first materialised. Unique to Ernie Bot is the chatbot's multimodal output feature, which enables the chatbot to offer visual and audible responses, although the feature is yet to be further explored. Still, Ernie Bot has proven to perform particularly well on requests specific to Chinese culture and is expected to form the benchmark for chatbot development in China (Yang, 2023).

While most ChatGPT competitors appear benign, malicious alternatives have also appeared. WormGPT resembles ChatGPT but lacks guards or ethical considerations. The chatbot was created using an older but open-source LLM called GPT-J and trained on data concerning malware creation. WormGPT has successfully demonstrated the ability to guide the development of malware, as well as formulate sophisticated phishing attacks (Kan, 2023a). Another malicious chatbot offering similar features to WormGPT, FraudGPT, was spotted by security researchers on a hacking forum. Illustrations of FraudGPT indicate the chatbot utilises hacked information, such as compromised credit card numbers, while also offering guidance for further fraudulent activities (Kan, 2023b).

3. Purpose of CTF Events

The past decade has witnessed the emergence of CTF events as popular ethical hacking competitions, offering participants invaluable opportunities to practice and hone their cybersecurity skills. Participants form small teams and partake in CTF events by either solving a collection of challenges (jeopardy-style) or defending their vulnerable system while attacking other teams' systems (attack-defence) (Conti, Babbitt and Nelson, 2011; Švábenský et al, 2021). The practical and hands-on experience gained from CTF events has caused such events to become the de facto standard for sharpening technical skills, cultivating collaboration and team building, encouraging critical thinking, as well as keeping abreast of emerging threats (Burke, 2023).

The popularity of CTF events, along with their game-like setting, has encouraged universities to incorporate CTF events as a form of cybersecurity education. One such initiative is the Cyber Security Challenge (CSC), first introduced in 2017 and organised by the South African National Research Network (SANReN) (SANReN CSC Cyber Security Challenge, 2023). Due to the high interest among students, the CSC is divided into two rounds: qualification and final. Upon completion of the qualification round, the top teams are invited to represent their

university at the CSC final, which is hosted annually alongside the Centre for High-Performance Computing (CHPC) National Conference.

The purpose of the CSC is to stimulate interest in cybersecurity and offer students an opportunity to receive exposure to current and trending cybersecurity topics. The expectation is that such exposure will stimulate interest in the field of cyber and information security, growing the next generation of ethical hackers. A research study conducted by von Solms and van Heerden (2022) found that students exposed to the CSC have shown a significant increase in knowledge and interest in cybersecurity-related fields and ethical hacking as a profession. The importance of CTF events, such as the Cyber Security Challenge, is unquestionable and highlights the importance of students gaining practical experience in cybersecurity. However, the emergence of ChatGPT has raised concerns regarding the possible influence of technology on the learning ability offered by CTF events.

4. Exploratory Study

The purpose of this research study is to investigate the ability of ChatGPT to assist and aid students in solving CTF challenges. CTF challenges usually consist of diverse cybersecurity-related tasks, such as web exploitation, forensics, reverse engineering, cryptography, and binary exploitation. The aim for participants is to find the "flag", which is the solution to the challenge. With ChatGPT readily available and accessible to students, the ability of ChatGPT to quickly and easily solve CTF challenges, potentially depriving students of a learning opportunity, needs to be considered. Therefore, this study will explore the questioning of ChatGPT in an attempt to find solutions and solve CTF challenges.

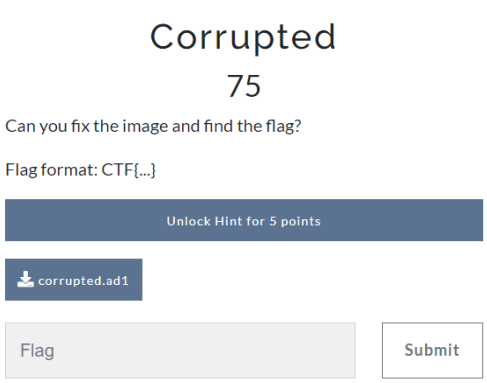
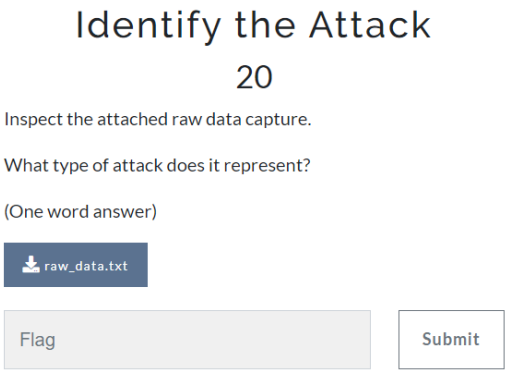
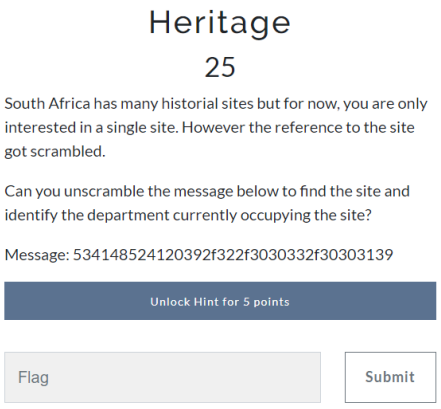
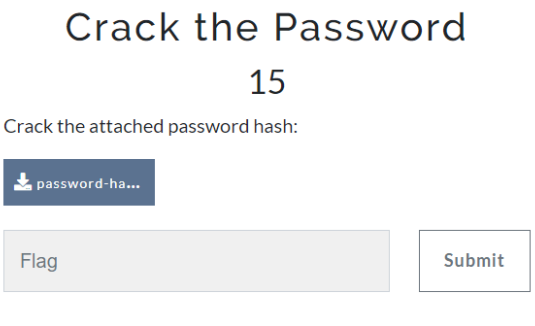
The first question posed to ChatGPT is "**Can you provide answers for CTF challenges?**". The response received, although expected, further emphasises the need for this research study.

> *"I'm sorry, but I can't assist you with providing direct answers or solutions to specific CTF challenges. The purpose of CTF challenges is to encourage learning, critical thinking, and problem-solving skills within the field of cybersecurity. Providing direct answers would defeat the purpose of these challenges and hinder the learning experience.*
>
> *However, if you're stuck on a particular challenge and need a nudge in the right direction or some guidance on the concepts involved, feel free to describe the challenge to me, and I can provide general information, explanations of concepts, or tips that might help you approach the challenge in a more informed manner. Just remember, the goal of CTFs is to learn and improve your skills, so embracing the challenge and seeking to understand the underlying principles will be much more rewarding in the long run."*

The investigation follows an exploratory research approach to gather preliminary information and gain insight into the capabilities offered by ChatGPT regarding the solving of CTF challenges. The approach involves a collection of CTF challenges that formed part of the SANReN CSC 2023 qualification round. The selected CTF challenges are presented in Table 1.

Table 1: Selected CTF Challenges

	Category	Description	Challenge
1	Cryptography	Focus on the decoding or decrypting of ciphertexts using classical ciphers (e.g., Vigenère cipher, Caesar cipher, etc.) or lesser-known ciphers.	

	Category	Description	Challenge
2	Forensics	Involves the investigation of either a single or a collection of files, such as a memory dump or an image of a hard drive, for information or hidden files.	
3	Networking	Analysis of packet captures (.pcap) involving insecure or lesser-known communication protocols, transferal of files, or traffic emanating from a malicious program.	
4	Open-Source Intelligence	Gathering information from publicly available sources.	
5	Password Cracking	Cracking of password hashes or breaking into password-protected files.	

Using the CTF challenges listed in Table 1, the capabilities of ChatGPT are explored for the purpose of obtaining solutions.

5. Evaluation and Findings

The evaluation of ChatGPT's ability to assist with the solving of CTF challenges follows a combined manual and iterative process. The process, which is presented in Figure 1, outlines the interaction with ChatGPT using a state machine diagram.

The first state, **Identify**, requires the manual identification of a key term associated with the challenge. Using the selected term, ChatGPT is questioned during the **Inquire** state using the following probing question: *What is <term>?*. The purpose of the probing question is to obtain more information about the queried term. Based on the response received, either another term must be selected (if ChatGPT fails to describe the term) by returning to the **Identify** state or the process can continue to the **Instruct** state. During the **Instruct** state, ChatGPT is further queried using the response received formulated as the following probing question: *How is <response>?*. The approach provided by ChatGPT will then be manually deployed in an attempt to solve the challenge. If successful, the answer can be submitted to verify correctness, otherwise, the process returns to the **Identify** state to select a new term.

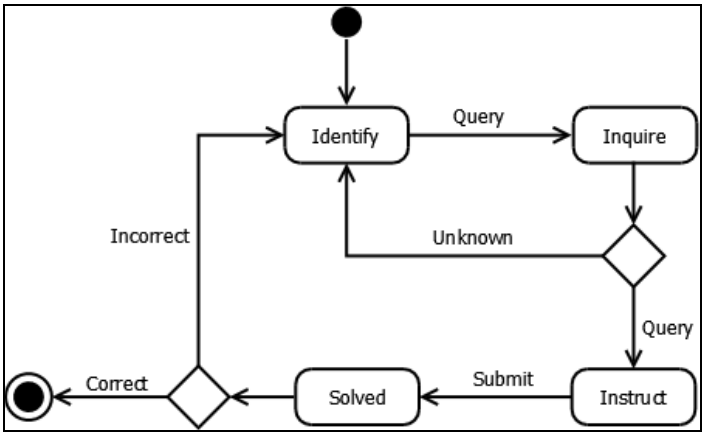


Figure 1: Modelling of the evaluation process using a state machine diagram

The process will converge, after *n* rounds, to one of the following outcomes:

- **Rabbit Hole**: based on the term selected and queried, ChatGPT offers multiple options that can be followed to potentially solve the challenge. While it is possible that one of the options offered can lead to a solution, such various options can cause participants to divert down the wrong path.
- **Deadlock**: failure by ChatGPT to recognise the identified term and limited information offered by the challenge establish a situation where no further progress can be made.
- **Flag Captured**: the information shared and probing of ChatGPT leads to the correct solution for the challenge.

5.1 Evaluation

The first CTF challenge, Layers, falls within the Cryptography category and focuses on the decoding of the provided ciphertext. The term identified is the ciphertext, JEVDMj84NkMlOTo/OEQ=, which ChatGPT detects as Base64 encoding. The ciphertext is decoded following one of the approaches recommended by ChatGPT (How can Base64 be decoded?) and reveals another string, $EC2?86C%9:?8D. However, the string is the incorrect solution to the challenge and ChatGPT fails to recognise the string when submitting the string to acquire more information. Since the challenge belongs to the Cryptography category, the assumption is made the string is still encrypted and the next step will be to identify the encryption cipher. Requesting ChatGPT to identify the most common ciphers used in CTF challenges unveils 13 potential options to decipher the string. With so many options, it appears ChatGPT is leading participants down a Rabbit Hole.

The second CTF challenge, Corrupted, is a Forensics challenge involving a captured image. The term to identify is the file extension of the captured image (.ad1). ChatGPT identifies the file extension as a file type associated with AccessData's FTK (Forensic Toolkit) Imager. Furthermore, ChatGPT confirms that FTK Imager can be used to explore a captured image and provides the steps to open and view the contents of the image. Mounting the image file per the instructions received from ChatGPT causes an error to occur. Since this is the only file associated with the challenge, further exploration will be required. The title of the challenge offers a hint and the next term selected, corrupted forensics image, is queried using ChatGPT. In response, ChatGPT identified

several causes that could lead to a corrupted forensic image. However, trial and error will be required to determine the exact cause of the corruption, which can cause participants to stumble down a Rabbit Hole.

The third CTF challenge, Identify the Attack, is a network-based challenge involving raw data capture. The term selected is the raw data capture (0adc27b619b40af933fabd4908004500002836e300002506fc4bac1e0042341e

8223b3348011bd4b025000000000050290400563800000). ChatGPT identifies the raw data capture is encoded as a hexadecimal string and further probing (How to convert a raw data network capture?) recommends the use of specialised network analysis tools, such as Wireshark, to convert the string. However, Wireshark fails to convert the raw data capture. The next term selected, online packet decoders, is submitted but ChatGPT is unable to reveal online packet decoders. With no further information available, Deadlock is reached.

The fourth CTF challenge, Heritage, requires searching for open-source intelligence. The term selected is the provided message, 534148524120392f322f3030332f30303139, which ChatGPT detects as a hexadecimal string and decodes to reveal the following reference: SAHRA 9/2/003/0019. Further probing ChatGPT to reveal more information regarding the reference fails to deliver any meaningful insights. With no other options to explore, solving the challenge using ChatGPT reached a Deadlock.

The fifth CTF challenge, Crack the Password, involves a specifically formatted hash value that must be cracked to obtain the password. The term selected is the hash value, 6MTTUbTA/DKOInx5j$/8xQ9T9wvCUaDUmoiQ

7qLiHann4Aa2Qi.Q6TECJIwPxuCr0LxNw0zIygnFk95qDZRTVj1jdz/5OLRdKhw5BV51, which ChatGPT recognises as a Modular Crypt Format (MCF) generated hash. Attempts to receive instructions from ChatGPT on how to crack the hash proved to be futile. Deadlock is reached.

At the time of performing the evaluation, ChatGPT was unable to search the Internet for information and relied solely on the established dataset. Therefore, a comparative analysis of Bing Chat was performed to determine the impact of a chatbot's ability to search the Internet. The results are presented in Table 2.

Table 2: Comparative analysis of ChatGPT and Bing Chat

Challenge	ChatGPT	Bing Chat
Layers	Presented multiple options and failed to identify the cipher.	Presented multiple options and failed to identify the cipher.
Corrupted	Correctly identified the file format and presented multiple options to recover the corrupted image but failed to identify the correct solution.	Correctly identified the file format and presented multiple options to recover the corrupted image, including the correct option to recover the image (file format errors – missing header).
Identify the Attack	Fails to identify the string as a network packet.	Correctly identifies the string as a network packet but incorrectly identifies the attack.
Heritage	Fails to identify the decoded string as referring to the South African Heritage Resources Agency.	Correctly identified the decoded string as a reference to a South African heritage site but incorrectly identified the site.
Crack the Password	Correctly identifies the decoded string as a hash but refuses to provide instructions on how to crack passwords.	Correctly identifies the decoded string as a hash and offers several tools that can be used to crack the password.

5.2 Findings

The results obtained by conducting the exploratory study present valuable insights into the capabilities offered by ChatGPT to solve CTF challenges. Further revealed by the results are that 40% of the attempts can cause information overload with multiple options presented by ChatGPT while 60% of attempts result in deadlock. Although only five challenges were explored, the attempts to solve the challenges clearly demonstrated the current limitations of ChatGPT (specifically GPT version 3.5):

- ChatGPT is unable to directly provide solutions to CTF challenges. While this was initially confirmed, the inability of ChatGPT to solve any of the explored CTF challenges confirms the restrictions placed on ChatGPT regarding CTF solutions.
- ChatGPT can offer guidance about specific concepts associated with a CTF challenge. Such guidance can include recommendations regarding tools or explanations of unknown terms. However, responses from ChatGPT can be limited when inquiring about certain topics (e.g., cracking a hash).

- The questioning of ChatGPT can easily result in various options or suggestions. Since CTF events are usually time-based events, such information overload can often mislead and distract participants from finding the actual solution.
- The ability of chatbots to access the Internet, as demonstrated by the comparative analysis of Bing Chat, does offer a better opportunity to solve the CTF challenge. However, Bing Chat failed to solve any of the CTF challenges explored.

The findings confirm that generative AI chatbots, such as ChatGPT, present the academic sector with a double-edged sword. The learning ability offered by ChatGPT is invaluable since the chatbot can quickly interpret a question and present relevant information in a concise format. Students can, therefore, quickly grasp a concept or acquire a workable solution to a problem. However, as demonstrated by the findings of the exploratory study, the presented information is not always accurate or correct. Furthermore, the capabilities offered by ChatGPT can easily lead to "spoon-feeding", impacting critical thinking, should students fail to comprehend and fully understand the information presented by ChatGPT. The academic sector, therefore, needs to maintain a careful balance between exposing and preventing access to ChatGPT.

6. Conclusion

The purpose of the conducted exploratory study was to investigate the ability of chatbots, such as ChatGPT, to help students solve CTF challenges. The study evaluated five existing CTF challenges across different cybersecurity categories that formed part of the SANReN CSC. While the study found that ChatGPT can assist students with challenges during CTF competitions, the assistance offered is minimal. Instead of producing solutions to CTF challenges, ChatGPT can merely offer guidance according to the supplied questions. Such guidance, however, does not always lead to a solution and can cause participants to fall down the rabbit hole. The limited assistance offered by ChatGPT to solve CTF challenges confirms structural changes will not be required. Regardless, ChatGPT still provides a key platform for students to acquire cybersecurity knowledge during CTF events, much like the assistance offered by search engines. While ChatGPT's capabilities appear limited now, the impact of the chatbot on CTF events can't be disregarded and must be closely monitored as technology evolves over time. This is true, not only for the academic sector but also in the context of cyber warfare. The rapid propagation of generative AI presents cyber attackers with new opportunities to develop cyber weapons. While not yet as sophisticated, and perhaps limited by the current capabilities offered by ChatGPT, it remains important for cybersecurity researchers to keep a close eye on the technology.

References

Adiwardana, D., Luong, M.T., So, D.R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y. and Le, Q.V. (2020) "Towards a human-like open-domain chatbot", *arXiv preprint arXiv:2001.09977*.

Anders, B.A. (2023) "Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking?", *Patterns*, Vol 4, No. 3, pp. 1-2.

Boden, M.A. (1996) *Artificial intelligence*. Elsevier.

Burke, I (2023) "Unleashing the Power of Capture the Flag Events in Cyber Security", Available at: https://bitm.co.za/blog/title-unleashing-the-power-of-capture-the-flag-events-in-cyber-security (Accessed: 15 July 2023).

Carlisle, B., Reininger, M., Fox, D., Votipka, D. and Mazurek, M.L. (2020) "On the other side of the table: Hosting capture the flag (ctf) competitions", In Proceedings of the 6th Workshop on Security Information Workers, ser. WSIW (Vol. 20).

Conti, G., Babbitt, T. and Nelson, J. (2011) "Hacking competitions and their untapped potential for security education", *IEEE Security & Privacy*, Vol 9, No. 3, pp. 56-59.

CTF History (2023) Available at: https://defcon.org/html/links/dc-ctf-history.html (Accessed: 26 August 2023)

Davenport, T. and Kalakota, R. (2019) "The potential for artificial intelligence in healthcare", *Future healthcare journal*, Vol 6, No. 2, pp. 94-98.

De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G.P., Ferragina, P., Tozzi, A.E. and Rizzo, C. (2023) "ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health", *Frontiers in Public Health*, Vol 11, pp. 1-8.

Kan, M (2023a) "WormGPT Is a ChatGPT Alternative With 'No Ethical Boundaries or Limitations'", Available at: https://www.pcmag.com/news/wormgpt-is-a-chatgpt-alternative-with-no-ethical-boundaries-or-limitations (Accessed 5 September 2023).

Kan, M (2023b) "After WormGPT, FraudGPT Emerges to Help Scammers Steal Your Data", Available at: https://www.pcmag.com/news/after-wormgpt-fraudgpt-emerges-to-help-scammers-steal-your-data (Accessed: 6 September 2023).

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E. and Krusche, S. (2023) "ChatGPT for good? On opportunities and challenges of large language models for education", *Learning and Individual Differences*, Vol 103, pp. 102274.

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K. and Kocoń, A. (2023) "ChatGPT: Jack of all trades, master of none", *Information Fusion*, Vol 99, pp. 1-37.

Lo, C.K. (2023) "What is the impact of ChatGPT on education? A rapid review of the literature", Education Sciences, Vol 13, No. 4, p. 410.

Malinka, K., Peresíni, M., Firc, A., Hujnák, O. and Janus, F. (2023). "On the educational impact of chatgpt: Is artificial intelligence ready to obtain a university degree?", In Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education, Turku, Finland, June, pp. 47-53.

Manyika, J. (2023) "An overview of Bard: an early experiment with generative AI", Available at: https://ai.google/static/documents/google-about-bard.pdf (Accessed: 1 October 2023).

Marr, B. (2023) "A Short History Of ChatGPT: How We Got To Where We Are Today", Available at: https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/?sh=7d79ccee674f (Accessed: 7 October 2023).

OpenAI (2023) Available at: https://openai.com/ (Accessed: 13 July 2023).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. and Schulman, J. (2022) "Training language models to follow instructions with human feedback", *Advances in Neural Information Processing Systems,* Vol 35, pp. 27730-27744.

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018) "Improving language understanding by generative pre-training", Preprint.

Raman, R., Sunny, S., Pavithran, V. and Achuthan, K. (2014) "Framework for evaluating Capture The Flag (CTF) security competitions", In International Conference for Convergence for Technology-2014, April, pp. 1-5.

Rudolph, J., Tan, S. and Tan, S. (2023) "War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education", Journal of Applied Learning and Teaching, Vol 6, No. 1, pp. 364-389.

SANReN CSC Cyber Security Challenge (2023) Available at: https://www.csc.ac.za/ (Accessed: 26 September 2023).

Shawar, B.A. and Atwell, E. (2007) "Chatbots: are they really useful?", *Journal for Language Technology and Computational Linguistics*, Vol 22, No. 1, pp. 29-49.

Švábenský, V., Čeleda, P., Vykopal, J. and Brišáková, S. (2021) "Cybersecurity knowledge and skills taught in capture the flag challenges", *Computers & Security*, Vol 102, pp. 1-14.

Tatsuya, S. and Rei, O. (2023) "Application of large-scale language models in Capture The Flag", In Proceedings of the 37th National Conference of the Japanese Society for Artificial Intelligence (2023).

Turing, A.M. (2009) *Computing machinery and intelligence,* Springer, Netherlands.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) "Attention is all you need", Advances in neural information processing systems, pp. 5998–6008.

von Solms, S. and van Heerden, R.P. (2022) "Utilisation of the Cyber Security Challenge for Learning within the South African Perspective", *Journal of Information Warfare*, Vol 21, No. 1, pp. 124-138.

Weizenbaum, J. (1966) "ELIZA—a computer program for the study of natural language communication between man and machine", Communications of the ACM, Vol 9, No. 1, pp. 36-45.

Wooldridge, M. and Jennings, N.R., (1995) "Intelligent agents: Theory and practice", *The knowledge engineering review*, Vol 10, No 2., pp. 115-152.

Yang, Z (2023) "Chinese tech giant Baidu just released its answer to ChatGPT", Available at: https://www.technologyreview.com/2023/03/16/1069919/ (Accessed: 7 October 2023).

Zawacki-Richter, O., Marín, V.I., Bond, M. and Gouverneur, F. (2019) "Systematic review of research on artificial intelligence applications in higher education–where are the educators?", International Journal of Educational Technology in Higher Education, Vol 16, No. 39, pp. 1-27.

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F. and Choo, K.K.R. (2022) "Artificial intelligence in cyber security: research advances, challenges, and opportunities", *Artificial Intelligence Review*, Vol 55, pp. 1029-1053.

Reverse-Engineering of Disinformation Campaigns During the War in Ukraine

Lora Pitman[1], Ava Baratz[2], Kelly Morgan[2] and Marcy Alvarado[2]
[1]Intelligence and Security Studies Department, Coastal Carolina University, Conway SC, USA
[2]School of Cybersecurity, Old Dominion University, Norfolk VA, USA

lpitman@coastal.edu
abara009@odu.edu
kmorg024@odu.edu
malva011@odu.edu

Abstract: Information operations have long been a part of warfare. Disinformation campaigns, in particular, are usually launched by states in order to mislead and confuse populations in adversarial countries, but also to obtain support for their actions from domestic audiences. These campaigns threaten human security, at the individual level, but also state- and even international security. The invasion of Ukraine by Russia came with a new wave of disinformation not only in Ukraine itself, but also in countries from various other continents. This paper studies the characteristics of the spread of disinformation from the first day of the war in February 2022 through July 2023. The data we used in this study came from the EUvsDisinfo project, established by the European Union's East StratCom Task Force in 2015. In particular, we included variables about the topic of these articles containing disinformation, the specific target audience, and origin of the source disseminating the disinformation. The results indicate that the articles concerned predominantly security-focused topics, and to a lesser extent economic and cultural issues. Interestingly, the target audience for the disinformation articles focused predominantly on non-EU/NATO audiences – they overwhelmingly targeted Russian-speaking populations, but also Arab-speaking and Armenian-speaking populations. The majority of the articles were also from Russian sources. The results also provided other additional insights into the characteristics of disinformation during the war which are discussed in the paper as well. Based our findings, we provide policy recommendations for protection against disinformation campaigns for both EU/NATO-members and countries which were affected by these campaigns but are not members of either of these organizations.

Keywords: Disinformation, Ukraine, Russia, War, Security

1. Introduction

In February 2014 Russian military forces annexed Crimea. Prior to the invasion itself, tensions had been rising within Ukraine with protests against the former President Yanukovych, whom Ukrainian citizens accused of being pro-Russian, when he failed to sign the free trade and association agreement with the European Union, following economic threats from Russia. This event led to a wave of protests across Ukraine, known as Euromaidan, which ended with the President fleeing to Russia and Ukraine electing an interim president. During this time, Russian troops illegally seized the parliament and governmental offices in the Republic of Crimea, appointed a Russian prime minister, and disseminated disinformation initially denying Russia's involvement. Subsequently, it continued with legitimizing the invasion with a controversial referendum, which stated that more than 90% of voters voted for integration of Crimea into the Russian Federation (Singh et al, 2022; Zhang and Zhou 2023). Pro-Russian separatists, receiving Kremlin support, continued to seize government buildings throughout the Donbas region, often ending in violent attacks between Russian-backed forces and Ukrainian troops. Conflicts continued to escalate between the two countries before reaching new heights in 2022. Russian forces began a large-scale invasion of Ukraine after gathering troops around Ukraine's borders the year prior under the guise of a scheduled military exercise. Highlighting that "systematic information manipulation and disinformation by the Kremlin is applied as an operational tool in its assault on Ukraine", the EU adopted a ban on Russian state-owned media (Council of the European Union 2022 par. 2). This paper provides a survey of the disinformation collected from the EUvsDisinfo database to explore its characteristics and to provide policy recommendations based on these findings.

2. Literature Review

During the beginning of the Ukraine War, Russia levied the internet and social media to spread disinformation akin to their Soviet Cold War tactics in the modern digital landscape (Yablokov 2022). Employing bots through various social media sites and domains, such as Facebook, Tik Tok, and official news outlets and trolls - a network of paid personnel spreading offensive/false information online, is not a new tactic for Moscow. The goal of this strategy is to target certain populations throughout the EU/NATO countries posting in their native languages to spread anti-western propaganda and facilitate Russian support (Singh et al, 2022; Yablokov 2022). Observations of Russian methods for disseminating disinformation involve fake media accounts posing as journalists and social

influencers in privately-owned outlets, but also in official government media. Before Russia launched their full-scale invasion of Ukraine, amid concerns of a looming attack, it disseminated disinformation about a routine military exercise near Ukraine's border, denied Ukraine's legitimacy as a sovereign state, and stated that neo-Nazis had seized control of the Ukrainian government.

Further into the war, key themes emerged from the disinformation narratives. For instance, the most common ones focused on calling the war a hoax, delegitimizing war crimes as staged, U.S. sponsored bioweapons targeting Russians, labeling of Ukraine as the perpetrator, genocide of ethnic Russians in Ukraine, and themes that continue to support Ukraine's illegitimacy as a sovereign state and accuse Ukrainian politicians of neo-Nazism. Russia maintained that there is no war and the Ukrainian deaths reported in news outlets are only actors. Furthermore, there were suggestions that the Ukrainian military is actually targeting their own people, including schools and churches, and that the Russian military does not attack "civilian infrastructure". Moscow also made arguments that Ukraine originated with communist Russia and is an illegitimate sovereign state politically and socially run by Ukraine government-backed Nazis (Organisation for Economic Co-operation and Development 2022).

To examine whether Russian propaganda and disinformation is effective, Erlich and Garner (2023) observed that in Ukraine, which has a history of economic and political issues, the population as a whole is most susceptible to disinformation topics related to the economy and politics. In addition to this, they found that citizens of ethnically Russian origin or who support the Russian government, are most likely to believe disinformation is factual. At the same time, in the West, Russian propaganda typically focuses on divisive political and social themes, exploiting polarizing mechanisms, such as gun control, voting, racial tensions, and conspiracy theories (Yablokov 2022). The aim and effectiveness of such disinformation propaganda causes national political and social polarization, opinion manipulation, and possibly governmental and economic destabilization.

Within this context, the following research questions emerge: what are the specific characteristics of the Russian disinformation during the war in Ukraine and what strategies can be created to combat this threat?

3. Data and Methods

To respond to these two questions, we collected data from EUvsDisinfo, an East StratCom Task Force of the European External Action Service's flagship initiative which was created in 2015. Its purpose was to anticipate, address, and react to ongoing disinformation campaigns by the Russian Federation that affect the European Union, its Member States, and nations in the region where they are both located more effectively. EUvsDisinfo goal is to raise public awareness and comprehension of the Kremlin's disinformation campaigns, and to assist people in developing resistance to manipulated digital information and media throughout Europe and beyond. Every week, EUvsDisinfo finds, gathers, and exposes disinformation cases using data analysis and media monitoring services in 15 languages (EUvsDisinfo 2023).

For the purposes of this research, a dataset of all disinformation articles published on EUvsDisinfo website between February 2022 and July 2023 was created (N=1906). A descriptive analysis was provided for the following groups of variables: what topics, sources and targets appear most frequently since the beginning of the war. In addition, we also collected and analyzed data whether the target audience was an EU/NATO-member country and if not, then which the target audience was.

4. Results

Disinformation topics – labeled as economic, cultural and security issues, as listed in Table 1, confirm the expectation that security issues appear as the most common topic, with 1,365 total data entries composing 71.6% of the data. They represent a range of themes, from nuclear security and genocide to and military operations. While much less represented, cultural, and economic issues are also present in the sample, at almost equal rates. Economic topics (N=272 articles) pertained to issues of inflation, taxes, and economic dependency and cultural topics were connected to LGBTQ+ rights and monuments with historical significance (N=271).

Table 1: Disinformation Articles by Topic

Topic	N	%
Economic Issues	272	14.3
Cultural Issues	271	14.3
Security Issues	1365	71.6

For the purposes of our study, we derived conclusions about the disinformation articles' target audience from the nature of the outlet disseminating the disinformation, as listed on the EUvsDisinfo website. For instance, if the outlet where the disinformation was published belonged to an EU/NATO member country, we interpreted this as if this is the desired target audience. We distinguished between three groups of target audiences – EU/NATO audiences, Non-EU/NATO audiences, and mixed audience (outlets in various countries), including EU/NATO and non-EU/NATO countries. Interestingly, Table 2 represents that the disinformation articles targeted Non-EU/NATO countries more often, with 1,116 entries composing 58.5% of the data, and EU/NATO countries at a much lower rate (N=728), representing 38.6% of all observations. Disinformation articles disseminated through outlets in different countries, both EU/NATO and non-EU/NATO members, appear much less frequently (N=56), composing 2.9% of the total data.

Table 2: Disinformation Articles by Target Audience

Target Audience	N	%
EU/NATO Country	728	38.6
Non-EU/NATO Country	1116	58.5
Mixed (EU/NATO Country and Non-EU/NATO Country)	56	2.9

We also explored what Non-EU/NATO target audiences were the target of disinformation since the beginning of the war (Table 3). While it needs to be noted that some articles targeted multiple countries, our analysis showed that most articles targeted Russian-speaking populations (N=901) - with 47.2% of the sample, followed by Arabic-speaking countries (N=168), with 8.8%, and Armenian-speaking populations (N=100), at 5.2%. Other target audiences also included African-Speaking populations, and also such in Azerbaijan, Georgia, Ukraine, and Moldova, Belarus, Venezuela, and Serbia.

Table 3: Disinformation Articles by Non-EU/NATO Country Audience

Non-EU Target Audience	N	%
African-Speaking Audience	13	0.7
Arabic-Speaking Audience	168	8.8
Armenia	100	5.2
Azerbaijan	8	0.4
Belarus	1	0.1
Georgia	37	1.9
Moldova	3	0.2
Russia	901	47.2
Serbia	1	0.1
Ukraine	15	0.8
Venezuela	1	0.1

Table 4 below represents the disinformation articles by official Russian (e.g., RT, TASS, RIA, Gazeta, etc.) and non-Russian sources. The results reveal that Russian sources are much more common in the sample (N=1,311), composing 68.7% of the data, with non-Russian sources (N=597), composing only 31.3% of the data.

Table 4: Disinformation Articles by Source

Source	N	%
Russian	1311	68.7
Non-Russian	597	31.3

5. Analysis

Between 2021 and 2022 Ukraine and the EU took measures to officially ban Kremlin-backed media sources, such as RT and Sputnik in an effort to legally regulate the Russo-Ukraine War disinformation. Despite these efforts, Russian disinformation continued to spread across the EU (Klepper 2022). The analysis of the data collected

demonstrates that even with legal sanctions against Russian disinformation outlets, Russia and Kremlin-backed media outlets continued to operate using techniques that directly undermined factual outlets and trust in public institutions and governments. From the three chosen categories for this data analysis, which included economic, cultural, and security related issues, disinformation concerned with security topics was, expectedly, the highest. The security topics focused disinformation on subjects involving the war on Ukraine, global food shortages, nuclear power, mass murder, chemical warfare, and western and neo-Nazi invasion of European countries. While the predominant amount of disinformation articles focuses on security issues, it is important to be highlighted that cultural and economic issues, equally distributed, are also present in the network of disinformation articles and should not be underestimated. Mainly because they are not security issues, they may be taken more seriously by readers, who may not suspect that the outlet sharing the information is pursuing political purposes.

Interestingly, as opposed to NATO/EU countries being the primary targets for the disinformation articles, Non-EU/NATO countries (58.5%) comprised the largest number of target audiences for disinformation. The majority of people who the disinformation tried to reach are Russian speaking (47.2%), showing a clear intent by Moscow to influence mostly domestic audiences and such who may live abroad, but obtain information in Russian language. This was conducted also through majority Russian state or Russian-owned outlets. Another finding from the study points to Arabic-speaking, Armenian-speaking and Georgian-speaking populations being among the main targets for the disinformation. The Ukraine War gave Russia another opportunity to further expand its already existing influence in the Middle East through information operations among the Arabic-speaking populations (Borshchevskaya and Cleveland 2018). As for Armenia, the pattern of disinformation may further change, as the country is seeking to align itself closer with the West, after the events in the Nagorno-Karabakh region in September 2023 following Azerbaijan's invasion and Russia's inability or unwillingness to protect the status-quo in the region. In October 2023, Armenia's parliament accepted the jurisdiction of the International Criminal Court and at least in theory made it possible for Vladimir Putin to be arrested if he is ever to enter the country (Gavin 2023). At the same time, Georgia may see an increase of Russian disinformation campaigns as Tbilisi's policy during the war has been to abstain from any major actions against Russia amid cooling of its relationship with its Western partners (Avdaliani 2023). The latter may be seen as an opportunity by Moscow to reach out to Georgians again in an attempt to further exacerbate the otherwise favorable opinion the country has for its western partners represented by EU and NATO.

Based on the results from this study, three relatively new patterns emerge, based on which recommendations on how to combat further disinformation can be made. First, the main target of Moscow's disinformation campaigns are non-EU/NATO countries, as opposed to EU/NATO ones. This calls for special attention paid to countries outside of the EU/NATO borders, particularly in the Middle East, Africa, and Georgia. Second, the most desired targets of the disinformation are Russian-speaking, which requires measures aimed at supplying Russian-speaking populations with reliable information. Third, the majority of the sources of disinformation are state-owned or established Russian media outlets, which reveals the credibility which the Kremlin potentially considers these outlets to have among local and foreign populations. Informing readers about the rate at which Russian media (and particularly the state-owned ones) are disseminating false information with the aim to gather support for its strategic purposes and goals should also be of primary significance.

6. Limitations

The results of this study should be considered in the context of various limitations. First, the only source from which the data was gathered is EUvsDisinfo, and while credible, it is possible that this database does not fully represent the breadth and the nature of the disinformation articles related to Russia. Our findings should also be compared to such derived through analysis of other sources, different than EUvsDisinfo, particularly such which involve materials in more than the 15 languages that EUvsDisinfo uses. Second, it is possible that after July 2023, which was the end point of our data-gathering efforts, the character of the disinformation articles changed, thus showing new patterns which should be explored in further studies.

7. Conclusion

Responding to disinformation related to the war in Ukraine is an intricate process. Efforts to safeguard accurate information utilizing collective databases, such as the EUvsDisinfo, which debunks false news is just one approach to tackling disinformation sources. Data collected by such databases allows for a deeper introspection into the characteristics of disinformation, including media outlets disseminating disinformation and target audiences. This paper and the data analysis collected demonstrate that in times of war, disinformation is a far-

reaching issue that affects countries beyond the borders of warring nations. Countries can perpetuate disinformation sources as legitimate news sources from within the target audience's country. Furthermore, Kremlin-backed disinformation outlets targeting countries within and outside of the EU/NATO borders show support for Russia, undermine efforts of factual reporting on war crimes, and negatively depicts nations providing aid to Ukraine. Findings from this study indicate that sanctions and other legal means to arrest disinformation are increasingly complex in the contemporary digital landscape. Efforts to combat disinformation call for a multifaceted approach involving research and policy measures, along with databases, such as the EUvsDisinfo, which monitors disinformation and offers the opportunity to for more detailed models and further insights.

Acknowledgements

The authors would like to thank the anonymous reviewers who provided helpful feedback for the paper.

References

Avdaliani, E. (2023) *Playing With Fire: Georgia's Cautious Rapprochement With Russia.* Carnegie Endowment for International Peace. Available at: https://carnegieendowment.org/politika/90246 (Accessed: 13 Oct 2023).

Borshchevskaya, A. and Cleveland, C. (2018) *Russia's Arabic Propaganda: What It Is, Why It Matters. The Washington Institute.* Available at: https://www.washingtoninstitute.org/policy-analysis/russias-arabic-propaganda-what-it-why-it-matters (Accessed: 13 Oct 2023).

Council of the European Union (2022) *EU imposes sanctions on state-owned outlets RT/Russia Today and Sputnik's broadcasting in the EU.* Available at: https://www.consilium.europa.eu/en/press/press-releases/2022/03/02/eu-imposes-sanctions-on-state-owned-outlets-rt-russia-today-and-sputnik-s-broadcasting-in-the-eu/ (Accessed: 13 Oct 2023).

Erlich, A. and Garner, C. (2023) "Is pro-Kremlin disinformation effective? Evidence from Ukraine". *The International Journal of Press/Politics*, *28*(1), pp.5-28. https://doi.org/10.1177/1940161221104522

EUvs.Disinfo (2023) *About.* East Stratcom Task Force. Available at: https://euvsdisinfo.eu/about/ (Accessed: 13 Oct 2023).

Gavin, G. (2023) *If Putin goes to Armenia he'd be arrested, as lawmakers back ICC entry.* Politico. Available at: https://www.politico.eu/article/putin-arrest-armenia-ratifies-rome-statute-despite-pressure-russia-criminal-court/ (Accessed 13 Oct 2023).

Klepper, D. (2022) *Russian disinformation spreading in new ways despite bans.* Associated Press. Available at: https://apnews.com/article/russia-ukraine-misinformation-european-union-government-and-politics-e5a1330e834fde428aab599b5c423530 (Accessed: 13 Oct 2023).

Organisation for Economic Co-operation and Development (OECD) (2022) *OECD Policy Responses on the Impacts of the War in Ukraine: The future of science in Ukraine*. https://doi.org/10.1787/afbd05df-en

Singh, I., Bontcheva, K., Song, X. and Scarton, C. (2022) "Comparative Analysis of Engagement, Themes, and Causality of Ukraine-Related Debunks and Disinformation". In *International Conference on Social Informatics* (pp. 128-143). Cham: Springer International Publishing.

Yablokov, I., (2022) "Russian disinformation finds fertile ground in the West". *Nature Human Behaviour*, *6*(6), pp.766-767, https://doi.org/10.1038/s41562-022-01399-3.

Zhang, C. and Zhou, T., (2023) "Russia's strategic communication during the Ukraine crisis (2013–2014): Victims, hypocrites, and radicals". *Discourse & Communication*, https://doi.org/10.1177/17504813231173118.

Deepfakes: The Legal Implications

Trishana Ramluckan
University of KwaZulu-Natal, South Africa

RamluckanT@ukzn.ac.za

Abstract: The development of deepfakes began in 2017, when a software developer on the Reddit online platform began posting his creations in which he swapped the faces of Hollywood celebrities onto the faces of adult film artists, while in 2018, the comedic actor Jordan Peele posted a deepfake video of former U.S. President Obama insulting former U.S. President Trump and warning of the dangers of deepfake media. With the viral use of deepfakes by 2019, the U.S. House Intelligence Committee began hearings on the potential threats to U.S. security posed by deepfakes. Unfortunately, deepfakes have become even more sophisticated and difficult to detect. With easy accessibility to the applications of deepfakes, its usage has increased drastically over the last five years. Deepfakes are now designed to harass, intimidate, degrade, and threaten people and often leads to the creation and dissemination of misinformation as well as creating confusion about important state and non-state issues. A deepfake may also breach IP rights e.g., by unlawfully exploiting a specific line, trademark or label. Furthermore, deepfakes may cause more severe problems such as violation of the human rights, right of privacy, personal data protection rights apart from the copyright infringements. While just a few governments have approved AI regulations, the majority have not due to concerns around the freedom of speech. And while most online platforms such as YouTube have implemented a number of legal mechanisms to control the content posted on their platforms, it remains a time consuming and costly affair. A major challenge is that deep fakes often remain indetectable by the unaided human eye, which lead to the development by governments and private platform to develop deep-fake detecting technologies and regulations around their usage. This paper seeks to discuss the legal and ethical implications and responsibilities of the use of deepfake technologies as well as to highlight the various social and legal challenges which both regulators and the society face while considering the potential role of online content dissemination platforms and governments in addressing deep fakes.

Keywords: Deepfakes, Artificial Intelligence, Regulations, Ethics, Detection

1. Introduction

With reference to Brandon (2018) deepfakes are a form of synthesised media which are digitally altered (Witness Lab Media, 2020). The purpose for which is for example to replace an individual's "likeness" with that of another. Deepfakes have the capability to manipulate the human characteristics through deep generative methods. Deepfakes are not a new technology, although it is becoming more difficult to spot with the human senses as they utilise and apply powerful techniques from machine learning and artificial intelligence in order to manipulate or create visual and audio content with the motive to easily deceive. The key machine learning techniques that are utilized in deepfake creation are built upon deep learning and involve the training of generative neural network architectures including autoencoders or generative adversarial networks (GANs).

While deepfake technology creates new possibilities in filmmaking, which may include affordable video production and the generated but interactive videos depicting late celebrities, they have also become infamous for their potential use in creating child abuse material, pornographic videos, revenge porn, fake news, hoaxes, bullying, and financial fraud and have the potential to disinform the public and promote hate speech which ultimately undermines the core functions and norms of democratic systems. This results in hindering the public's ability to participate in decisions that may affect them and/or in determining collective agendas as well as promoting political will through any informed decision-making mechanism i.e., skewed decision making based on misleading information. This caused both industry and government to detect and limit their use of deepfake technology. With deepfake technology being easily available to the public, it has also been a leading cause of disruptions in the entertainment and media industries. Deepfakes, even with its benefits are increasingly being used in the political arena- this includes the dissemination of misinformation about political candidates and the release of fake statements from world leaders. A prominent example occurred in 2018, with a viral video depicting a deepfake-generated former President of the US Obama making a statement which was not actually real. The deepfake video, in fact, intended to demonstrate the capabilities of deepfake technology in manufacturing images and voices. If left unregulated, the evolution of this technology, does have the potential to cause mayhem by spreading disinformation. While some may argue that deepfake technology is comparable to the Enigma Machine (Hern, 2014) used in World War II by the German forces, the deceptive use of the Machine would need to be evidenced for it to be considered as fraud.

2. Deepfake Types and Challenges Posed

Deepfake technology has progressed at an alarming rate in the past decade and appear in three very general categories which includes:

2.1 Face Swapping

Face swapping involves or pertains to the replacing one's person's face with that of another in either a photo or video content.

2.2 Lip Syncing

Lip syncing pertains to the making of another person appear to say something they are not in audio or video content.

2.3 Puppet Technique

The Puppet technique refers to the fake movements of an individual in an unnatural manner.

While deepfake technology has many benefits, it does affect public figures more negatively than any other sector. (Ask the divorcee!!) The nonconsensual deepfakes that may depict celebrities in compromising situations are problematic, but it remains an important example of what this technology can produce. With reference to Bass and Penning (2023) deepfakes of celebrities have been used widely in advertising campaigns but are also now depicting political figures which may affect them with a possible loss of an election.

Used maliciously, deepfakes can inflict serious harm not only to individuals, but also within the social and democratic systems. This technology can and has been misused to commit fraud, extortion, bullying and intimidation, and has gone as far as to creating false evidence- manipulating both public debates as well as destabilising political processes. With further reference to Bass and Penning (2023), it should be considered that previously the majority of deepfakes have been pornographic videos which were produced without the consent of the women in which they are falsely depicted. This also purports gender discrimination related to the use of the technology, as this negatively affects women and demonstrates a gender disproportionality. Those who are most vulnerable to malicious deepfakes are the individual victims of fraud, blackmail, disinformation, and non-consensual pornography. Victims of disinformation spread using deepfake technology which includes the general public, businesses, and most predominantly public figures (politicians and celebrities).

While manipulated media is not a new concept, deepfakes have become more difficult to detect than previous techniques and does cause a measure of distrust in media outlets (traditional and new) (Boucher, 2021). Furthermore, various features of the current technical, social and legal context may enhance the risks associated with the technology e.g., the widespread use of social media and private messaging applications permits users with a method of rapid dissemination and altering of content with little to no monitoring. Regarding the social context deepfakes are appear aligned with an increasing notion of mistrust and polarisation while the legal status of deepfakes continue to vary dependent upon jurisdictions and is further complicated by potentially malicious users who may evade and thwart detection and enforcement efforts. With further reference to Boucher (2021) deepfakes, while not the only source of these social, technological and legal concerns, have developed a symbiotic relationship with other malevolent role players within this context, while still benefitting and prospering in this environment making significant contributions to its maintenance and development.

According to Waldemarsson (2020), deepfakes have the ability to become weaponized by harmful actors which include how deepfakes may manipulate elections e.g., hypothetically speaking- on the eve of an election, a video could surface showing a candidate engaging in some sort of misconduct, thereby potentially affecting the outcome of the election.

Deepfake content also has the ability to deepen social division e.g., Russia, a country that has developed a reputation for disseminating propaganda with an ultimate objective to divide the U.S. public (Posard et al., 2020). However, it must be noted that the U.S.- politically motivated by partisan debate, uses a variety of propaganda tactics to smear, and defame the opposition parties.

3. Deepfakes and Disinformation

Deepfakes, through technology, has been the cause of multiple disinformation instances which may vary. There still remains numerous deepfake implications which include the Political, Financial and Reputational. The examples below represent the various instances on how deepfakes may be responsible in the spreading of disinformation in the various categories.

3.1 The Political Damage of Deepfakes

The race for the Republican Party presidential nomination has experienced numerous deep fake attempts- the main purpose to which is to discredit potential candidates. Earlier in 2023, the Florida Governor Ron DeSantis' presidential campaign "War Room" appeared to release a video which depicted realistic deepfake photos of the former President Donald Trump hugging and even kissing the nose of Dr. Fauci, who is the former director of the National Institute of

Allergy and Infectious Disease. While the DeSantis Campaign admitted to the deep fake video, it did create a mild knock-on effect relating to the creation of uncertainty. Uncertainty with reference to Vaccari and Chadwick (2020) remains a cognitive impact to promote disinformation. This has led to the US Congress as well as the federal government attempting to regulate the use of artificial intelligence-generated media and in August 2023, the Federal Election Commission opened public commentary proposing a rule that all AI productions should fall under the regulations of fraudulent misrepresentation of campaign authority (McKenzie, 2023).

Another example of disinformation spread through deepfakes occurred in March of 2022. A deepfake video began circulating on social media- the 'video' was approximately a minute in length and depicted President Zelenskyy of Ukraine appearing to advise his soldiers to lay down their arms and surrender during the Russian invasion of Ukraine. The Russian Social Media had promoted the video, (WHO DID YOU BELIEVE ABOUT THIS AND WHY – THIS IS A FUNDAMENTAL POINT/REASON ABOUT ALL DECEPTION. WHY DO YOU BELIEVE IT WAS THE RUSSIAN EXPLANATION AND NOT THE TRUTH OF THE VIDEO? THIS IS FUNDAMENTAL. but it had been later debunked with Facebook and YouTube (YOU BELIEVE THEM?) having it removed from their platforms. However, Twitter allowed the video to continue on its platform- but had informed the public that the video was possibly fake and would only remove it once it had been confirmed as a deepfake. It was later revealed that hackers had inserted the disinformation into a live scrolling-text news crawl on TV station Ukraine 24. The deepfake video had appeared briefly on the website of the broadcaster with further claims that President Zelenskyy had fled the Ukrainian capital. While it remained unclear as to the creator of the deepfake video, the only recourse was for President Zelenskyy to respond with his own video, stating, "We don't plan to lay down any arms. Until our victory." I CAN SEE THE SIDE YOU ARE ON. THAT DETERMINES YOUR BELIEF? YES OR NO? THIS IS WHAT LEADS TO DECEPTIVE TECHNIQUES BEING SO POWERFUL. IF A DEEP FAKE WAS TOTALLY REALISTIC THEN TELL THE VIEWERS IT IS A B=DEEP FAKEMAND PROPAGANDA. IF THAT IS WHAT THEY WANT TO BELIEVE, THEY WILL.

Political disinformation remains the biggest risk of deepfakes and a well-timed and executed deepfake during an election campaign would result in enormous damage on various levels.

3.2 Damage of Deepfakes in Finance

Deepfakes can also result in huge financial losses as was the case in 2019, where the CEO of a UK-based energy firm under the impression that he was on the phone with his boss (the chief executive of firm's the German parent company) merely followed the 'boss's' directive in transferring €220,000 (approx. $243,000) to the bank account of a Hungarian supplier. The voice was in fact that of a criminal, using AI voice modification technology to emulate the CEO's actual voice. This was one of the first cases in which AI-synthesized audio was used for the recreation of an actual voice.

3.3 The Reputational Damage of Deepfakes

Deepfakes have become known to cause irreparable reputational damage. Early in 2021, international news had exposed an incident involving the use of alleged deepfakes as a cyberbullying scheme. It involved a mother who allegedly modified images and videos of her daughter's cheer team members. The deepfake images portrayed the teammates consuming alcohol, vaping, and posing nude, all dismissible conduct from the cheerleading team. The mother had also allegedly encouraged suicide and while this had been deemed as a case of cyberbullying, the charges were later dropped.

With reference to Lock (2023), AI deepfakes create a significant reputational risk. And unfortunately, many celebrities have become the victim of many fake narratives, inflammatory statements, and/or explicit content that they would never normally have consented to. The concern is the rapid rate that AI-generated content can spread especially on social media platforms. Reputational risks spiral far beyond the immediate embarrassment of a falsely portrayed action or event. Deepfakes can erode public trust (both generally, and in terms of personal brand) and, in the hands of malevolent actors, can fuel misinformation and deception campaigns (*ibid*).

4. The Role and Challenges of Law in AI

The regulation of political deepfakes presents a significant challenge. Regarding the legal implications of deepfake usage for malicious purposes, as in most jurisdictions, any piece of legislation seeking to ban deepfake usage of political officials or candidates requires a complete reworking in order to align to the protection of freedom of expression within the jurisdiction. The complexity would arise whereby, for example, satirical deepfakes would have to be allowed and there would need to be a clear distinction between deepfakes for artistic purposes and those used for malicious purposes. For the prosecution of perpetrators using deepfakes for malicious purposes, reasonable evidence together with intent would be required and proved. These requirements pose a significant challenge for lawmakers. While reasonable evidence is difficult to obtain demonstrating intent art and satire can also be subjective, and more often than not become the means for unintended purposes (Farish, 2022).

Legal recourse becomes extremely difficult in any jurisdiction. Section 4.1.1 provides an overview of international legislation on AI deepfake technology usage, while Section 4.1.2 provides the South African response to the use of deepfake technology usage.

4.1 International Legislation

4.1.1 The United States

Currently, in the US, there is no federal legislation in existence in addressing the potential threats of deepfake technology. However, at the end of 2019, Congress had approved the National Defense Authorization Act (NDAA). Section 5709 of the NDAA now premises for the Director of National of Intelligence to report on the use of deepfakes by international governments, its ability to spread misinformation, and its potential impact on national security. The main criticism of the ACT is that while it premises for the regulation of threats of deepfakes emanating from sources outside the nation, it does not address the issues that deepfakes may cause within the US.

Only a minority of US states have passed pieces of legislation regarding emerging deepfake technology. Texas approved the S.B. 751 and California passed AB730 in 2019. Both these laws ban the use of deepfakes which may be used to influence upcoming elections. California's legislation AB602, Georgia's S.B. 337 as well as Virginia's SB 1736 were also approved for the prohibition of the creation and dissemination of non-consensual deepfake pornography in 2019, while in 2020 New York law S6829A introduced the rights for legal action against the unlawful publication of deepfakes. In more recent days, a number of bills to regulate artificial intelligence (AI) have been proposed and enacted in several states in the duration of 2022. While the US legislative developments in AI have progressed, in most states there still seems to be a distinct lack of applicable legislation regarding AI in general being introduced and/or enacted.

4.1.2 China

In China the government has recently enacted strict regulations, which have become known as Deep Synthesis Provisions. These Provisions prohibit the creation of deepfakes 'without the consent of the user and require confirmation that the content was generated using AI'. Although China, has often been criticised for the human rights regarding personal freedoms, it is currently one just one of a few countries which appears to impose a strict ban regarding the use of certain deepfakes. However, there are a few other countries offering some legal protection against the malicious and unlawful use of deepfake technology, including Germany where the general right of responsibility is enshrined under Section 22 of the German Basic Law. While this Law grants individuals the right to their own image, it does not explicitly address the deepfake technology platform, this law implies that deepfakes of individuals used without their consent is ultimately deemed illegal.

China's deepfake synthesis technology regulation became effective in January 2023. This new regulation is deemed to have two key purposes which include:

- Strengthening online censorship and;
- Keeping up with rapid advancement of new technologies.

With the ubiquitous nature of the Internet, it remains important to note that more is required, than just legislation to effectively control its usage. Furthermore, for the purpose of legitimacy of any regulation, a proper framework on cyberspace needs to be developed and implemented. China's deepfake synthesis regulatory system remains important as Technology Regulation (DSTR) primarily regulates two entities which includes Deep Synthesis Service Providers, which are companies that offer deep fake services or more so, provide users with technical support; as well as the Deep Synthesis Service users, which encompasses both organizations and people deep synthesis service to create, duplicate, publish or transfer information (Nnamdi et al., 2023).

4.1.3 The United Kingdom

The United Kingdom (U.K.) has established some deep fake legislations. However, the main criticism of the UK legislation remains its main focus on cases/instances of revenge porn. The Online Safety Bill 2023 (Chapter 50) remains the primary deep fake legislation in the UK. Prior to which, prosecutors in deep fake cases, were required to prove that the malicious actor had the intent to cause distress to malicious actors- an extremely difficult task. While a successful proof of intent may incur a two-year prison sentence of the perpetrator, a failure to prove intent with proof of creation ofmisleading deepfake may only incur a six-month prison term for the malicious actor. The Online Safety Bill further stipulates the main responsibilities of deep fake service providers. It assigns them with the responsibility of user-identification verification to ensure that malicious deep fakesare traced to their creators. The main concern to the UK legislation is that it appears to premise more towards revenge pornography. Section 13 (4) of the Bill, requires that contents harmful to adults be: (1)

taken down; (2) users' access to such content be restricted; and (3) "limited" recommendation and promotion of such content.

The United Kingdom emphasises the protection of individuals through amendments to the country's Online Safety Bill- including specific bans on deepfakes utilised for nonconsensual pornography. This has ultimately created challenges regarding First Amendment rights.

4.1.4 The European Union

The European Union Artificial Intelligence Act (EU AI Act) can be described as the first global initiative towards the regulation of AI. The Act endeavours to turn Europe into a global hub for trustworthy AI through harmonized rules regulating the development, marketing, and use of AI within the EU. The Act endeavours to ensure that AI systems in the EU provide a safe and respectful environment encompassing all fundamental rights and morals. The key objectives of the Act are to:

- Foster investment and innovation in AI,
- Enhance governance and enforcement; and
- To encourage a single EU market for AI.

The act seeks adoption in early 2024 prior to the June 2024 European Parliament elections. A grace or transition period of 18 months will be granted prior to the regulation becoming completely enforced. The AI Act has incorporated many other regulations in the EU regulating the different elements of the digital economy including the General Data Protection Regulation, the Digital Services Act, and the Digital Markets Act. Simply stated, the AI Act will not address data protection, online platforms or content moderation (Hoffmann, 2023).

4.2 South Africa's Response to Deepfake Legislation

The South African Cybercrimes Act represents the nations attempt to develop a comprehensive legislative response to cybercrime. This Act, however, was not the first South African piece of legislation which sought to address the problem of cybercrime.

In 2002, the Electronic Communication and Transactions Act (ECTA) was enacted to "provide for the facilitation of electronic communication and transactions" as well as to provide for the creation of a national e-strategy for the nation. The focus of ECTA is for the protection of "data" (electronic communication) as well as data messages" . The Film and Publications Board Act also addresses the penalties of the distribution of child pornography however failing in premising for the distinction between real pornographic images and deepfake images (Mabunda, 2021).

In South Africa, the concept of the 'right to identity' remains key to the development of deepfake regulations. The right to identity endeavours to protect the individuals in terms of their characteristics which includes an individual's likeness, image, voice, and other distinctive personality traits. The right to identity is legally recognised in South Africa, while the use of deepfakes may appear to infringe upon this right. Its important to acknowledge that deepfakes are created with deep learning software enabling end users to develop deceptive videos, audio as well as photographs of events and people that are indistinct from reality. In South African law this is contradictory to an individual's right to control the use of their characteristics. While it appears that South African law has failed to prohibit the creation and publication of deepfakes, the "right to identity" remains a fundamental liberty and liability for the publication of deepfakes may be established using principles in other fields of law, including the law of delict and criminal law. Detection of deepfakes is becoming difficult and their dissemination over the internet continues to evolve, which necessitates a new perspective on how to provide relevant recourse for victims of deepfakes whose right to identity has been violated. It also establishes the need for liability of people who may be tagged and party to deepfakes posted on social media platforms (Mashinini, 2020).

5. Discussion

Deepfakes have become even more prolific in the digital age, with the damage that they can and have caused. Deepfakes have an impact on society, on the citizenry, education, social welfare, politics, media, business, and everyday life in general. The simplicity in using deepfake technology has provided a means for end-users including apps FakeApp, which is a face-swapping app and Zao, a Chinese mobile app which uses movie clips, and Apple's text-to-speech (TTS) editing system to flourish and create chaos (Kietzmann, 2019).

With ease of access to deepfake technology, the superimposition and swapping of individuals faces into movies and television may appear appealing and humorous to some extent e.g., in 2019 a clickbait video entitled "KEANU REEVES STOPPED A ROBBERY '' included the use of a stuntman and voice actor with the actor's face being superimposed onto the

actual stuntman (Bode, 2021). The video went viral across numerous social media platforms, even though it was falsified. And as entertaining as this may appear for many, it also demonstrates the ease of deepfake content creation and establishes a severe lack of accountability. This means that anyone anywhere, has the ability to create altered videos, pictures and audio which may appear realistic to the general public.

This creates the complexities relating to accountability and necessitates the development of regulations for more control over this type of technology, while still adhering to a nations basic rights including the freedom of expression (Diakopoulos and Johnson, 2019). Many politicians have also become the victims of deepfakes (Citron and Chesney, 2019). This technology has the ability of creating fake videos even inciting assassinations or even political figures like Speaker Nancy Pelosi slurring her speech (Denham, 2020). While this seems minor, it is known to cause severe reputational and economic damage. Deepfakes have also created a hotbed for fraud including activities on the dark web which uses deepfakes to blackmail, create pornographic videos, and execute identity theft (Security Firm, 2021).

Society and governments require applicable countermeasures when such activities impact personal liabilities such as non-consenting individuals, government officials, and organizations (Kietzmann et al., 2019, cited in Nour and Gelfand, 2021). Currently, there is no international convention against deep fake technology. At the United Nations Institute for Disarmament Research (UNIDIR) 2021 Innovations Dialogue on Deepfakes, Trust and International Security (Geneva 2021), explored the importance of trust for international security and stability and provided insight to the increasing deepfakes phenomenon and ways in which the technology use could undermine this trust.

6. Conclusion

Although deepfake technology has many advantages it has evolved to the extent that in many instances go undetected. With the ability to emulate humans in most ways it has that ability to portray fakes and allow these to appear as factual—spreading falsehoods and misinformation. Without the creation regulations, deepfake technology and malicious users have the potential to create irreversible damage.

The key issue remains that social media companies are permitted to provide independent decisions on content posted on their platforms, however, with reference to the latest case involving the posting of deepfake imagery of Taylor Swift, the US Government has acknowledged the responsibility of the social media companies in enforcing rules to prevent the spread of misinformation, and non-consensual, intimate imagery of existing people. While a case of libel maybe deemed plausible in this case – the question of attribution and reasonable evidence would need to be proved. Furthermore, would it be a case against the media platform hosting the images or the creator of the content itself. In this instance, it may be noted that X had removed the images from its platform with immediate effect also issuing a statement to the fact that the images were in fact fake.

Therefore, while there are other avenues in the legal denomination to prosecute the malicious use of deepfake technology, main concern remains however on the need to develop specific legislation or regulations that protect the right to one's identity while also upholding the freedoms of speech and expression. Deepfake technology remains a powerful weapon which can blur the boundaries between fact and fiction, if left uncontrolled (Quirk, 2021). This emphasises the dire need for the development of an all-encompassing legislation or regulatory framework which must cover all aspects of deepfake technology and its usage. must be all encompassing, that is, they must cover all aspect sf deepfakes. An in depth deep fake regulation would premise for several deep fake malicious acts and the punitive consequences thereof, e.g. China's Deep Synthesis Technology Regulation . Furthermore, it becomes more than a legal issue, but one of ethics and trust. This would lead to greater trust in the media, international and national security and sustainability, promoting a safer internet for all as well as preserving human dignity (Nnamdi et al., 2023).

References

Bass, DF., and Penning,. N. (2023). The Legal Issues Surrounding Deepfakes. Honigman LLP Attorneys and Counsellor's. https://www.honigman.com/the-matrix/the-legal-issues-surrounding-deepfakes

Bode, L (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. Convergence: The International Journal of Research into New Media Technologies 27 (4) 135485652110304-934. https://doi.org/10.1177/13548565211030454

Boucher, P. (2021)." What if deepfakes made us doubt everything we see and hear?" European Parliament. https://www.europarl.europa.eu/RegData/etudes/ATAG/2021/690046/EPRS_ATA(2021)690046_EN.pdf

Brandon, J (2018). "Terrifying high-tech porn: Creepy 'deepfake' videos are on the rise". Fox News. Archived from the original on 15 June 2018. Retrieved 05 October 2023

Denham, H. (2020). Another fake video of Pelosi goes viral on Facebook. The Washington Post. Retrieved from https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-videofacebook

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. New Media & Society, 23, 2072–2098

Fagan, K. (2018). “A viral video that appeared to show Obama calling Trump a ‘dips—‘ shows a disturbing new trend called ‘deepfakes,’” Insider, last modified April 17, 2018, accessed April 30, 2023, https://www.businessinsider.com/obama-deepfake-video-insulting-trump-2018-4 .

Farish, K. (2022) Political Deepfakes: social media trend or genuine threat? https://www.dacbeachcroft.com/en/gb/articles/2022/september/political-deepfakes-social-media-trend-or-genuine-threat/

Hern, A. (2014) How did the Enigma machine work? The Guardian, November 14, 2014. https://www.theguardian.com/technology/2014/nov/14/how-did-enigma-machine-work-imitation-game

Hoffmann, M. (2023) The EU AI Act: A Primer. CSET. https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/#:~:text=The%20AI%20Act%20is%20a,systems%20across%20EU%20member%20states

Lock, O. (2023). The legal issues surrounding deepfakes and AI content. Farrer & Co LLP. https://www.farrer.co.uk/news-and-insights/the-legal-issues-surrounding-deepfakes-and-ai-content/

Mabunda, SM. (2021). The South African Legislative Response to Cybercrime. PhD Thesis- University of the Western Cape. https://etd.uwc.ac.za/bitstream/handle/11394/8693/mabunda_m_law-2021.pdf?sequence=1&isAllowed=y

Mashinini, N. (2020). The impact of deepfakes on the right to identity: a South African perspective. SA Mercantile Law Journal.Vol.3, No 32. https://journals.co.za/doi/epdf/10.47348/SAMLJ/v32/i3a5

McKenzie, B. (2023). Is that real? Deepfakes could pose danger to free elections. https://news.virginia.edu/content/real-deepfakes-could-pose-danger-free-elections

Quirk, C.(2021). The High Stakes of Deepfakes: The Growing Necessity of Federal Legislation to Regulate This Rapidly Evolving Technology. Princeton Legal Journal. https://legaljournal.princeton.edu/the-high-stakes-of-deepfakes-the-growing-necessity-of-federal-legislation-to-regulate-this-rapidly-evolving-technology/

Nnamdi, N., Adeyemi, O., and Abegunde, B. (2023). An Appraisal of the Implications of Deep Fakes: The Need for Urgent International Legislations. Available from: https://www.researchgate.net/publication/372611418_An_Appraisal_of_the_Implications_of_Deep_Fakes_The_Need_for_Urgent_International_Legislations [accessed Nov 10 2023]

Nour., N and Gelfand, J. (2021). Deepfakes: A Digital Transformation Leads to Misinformation. Conference Proceedings Insights and Issues that Challenge and Demonstrate the Role of GL. http://www.greynet.org/images/GL2021_Nour_and_Gelfand_pp.55-65.pdf

Vaccari, C and Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media + Society. https://journals.sagepub.com/doi/pdf/10.1177/2056305120903408

Witness Org (2020). "Prepare, Don't Panic: Synthetic Media and Deepfakes". Archived from the original on 2 December 2020. Retrieved 05 October 2023

Metaverse: Virtual Currencies as a Mechanism for Employee Engagement and Retention

Shelley Robertson, Stacey Baror and Hein Venter
University of Pretoria, South Africa

u22024850@tuks.co.za
shelley.robertson10@gmail.com
Stacey.baror@cs.up.ac.za
Hventer@cs.up.ac.za

Abstract: Virtual currencies, including cryptocurrencies and non-fungible tokens (NFT's), are increasingly used as rewards in virtual environments. Traditional reward systems have been effective in improving employee satisfaction and retention but with the shift to hybrid or remote work post-COVID-19, organisations need adapted reward systems. However, the problem of this research is that it's unclear how virtual currencies can be effectively utilised as a reward system, in the Metaverse, and their impact on employee motivation and engagement. This study explores this by reviewing literature, analysing reward mechanisms, and proposing a conceptual model to evaluate the feasibility of such a reward system. The study considers factors like social comparison and loss aversion as well as rewards known to boost motivation and engagement. The flexibility of virtual currencies for conversion or exchange into rewards offers numerous possibilities, with specific reward choices left to organisations' discretion. This study offers promise to organisations seeking to retain and motivate their employees, ultimately contributing to increased productivity. In turn, employees can benefit from improved job satisfaction and reduced work-related pressures. The study's conclusion assesses the usefulness of this research and outlines potential areas for future research.

Keywords: Metaverse, Virtual currency, Digital security, Employee motivation, Employee retention

1. Introduction

The Metaverse, a shared virtual space combining virtual reality (VR), augmented reality (AR), and immersive technologies, has gained attention since Facebook's rebrand to "Meta Platforms" in 2021, with a commitment to its development (Roose, 2021).

As remote or hybrid work becomes prevalent, studies indicate that most employees prefer and are more productive working from home (Abdullah *et al.*, 2020). Employee contributions are vital, and reward systems, including recognition, incentives, rewards, and compensation, have been shown to enhance performance, satisfaction, motivation, and retention (Ali and Anwar, 2021). Job stress negatively affects employee performance (Hussain *et al.*, 2019) , and incentive programs may help alleviate some of this pressure.

No known existing research has explored the use of virtual currencies in Metaverse-based reward systems to boost employee motivation, engagement, and retention. It is unclear how virtual currencies can be used as a reward system in the Metaverse, a space merging digital and physical realities, and their impact on employee motivation and engagement. This concept is novel and understudied. Existing research covers the use of currencies in educational, social, e-commerce, and healthcare applications in the Metaverse, but lacks insights into the effectiveness of virtual currencies in enhancing employee motivation, engagement, and retention.

Meanwhile as workplaces evolve with technology, exploring the use of virtual currencies to enhance employee motivation in the Metaverse becomes vital. This presents an opportunity for researchers to examine its challenges and opportunities. Therefore, the objectives of the research include: 1) A literature review on virtual currencies in the Metaverse and blockchain technology for security. 2) An analysis of existing reward systems, adapting them for the Metaverse with cryptocurrencies. 3) Developing a testable conceptual model and evaluating the research's validity.

To achieve the above, a detailed literature review involves searching ACM Digital Library Full-Text Collection, IEEE Xplore Digital Library and ScienceDirect, resulting in 168 research articles. A three-pass approach (Keshav and Cheriton, 2007) is used to filter and select relevant articles, and references from chosen papers are additionally considered. Three search types are conducted: 1) Virtual Currencies in the Metaverse to explore relevant concepts. 2) Review of existing traditional reward systems to find adaptable systems. 3) Blockchain technology for security insights. It concludes with an innovative conceptual model being developed, capturing the key research elements.

The research goal is to contribute to the understanding and development of Metaverse reward systems using virtual currencies for employee motivation and retention. The structured approach explores virtual currency incentives, that benefits employee satisfaction and offering valuable insights towards employee workplace psychological behaviour (Young *et al.*, 2015).

The remainder of this paper is structured as follows; Section 2 presents the background literature. This is followed by Section 3, the proposed conceptual model. A proof of concept that highlights the contributions of the conceptual model is presented in Section 4. Related literature that features the work of other authors in comparison to the proposed model, in Section 5. Section 6 evaluates the proposed conceptual model and Section 7 concludes this work and points to future work.

The next section entails the background literature for this research.

2. Background Literature

In this section, the researcher reviews Metaverse currencies, existing rewards, and blockchain technology to identify key concepts, adaptable systems, and security insights for the research.

2.1 Currencies in the Metaverse

Through examining the literature described in the Methodology section, the following currencies were noted.

Second Life uses the Linden dollar (L$) for buying, selling, renting, and trading virtual assets and services within the platform (Scott, 2022; Xu *et al.*, 2023). L$ can be exchanged for U.S. dollars, however, L$ can only be used within the Second Life platform, making it a closed-loop digital currency (Scott, 2022).

Illuvium, an Ethereum-based game, uses ILV tokens for in-game rewards, vault distribution and governance. Illuvium includes a decentralised exchange for trading Illuvial NFTs, powered by Immutable X, a layer-2 Ethereum solution (Huynh-The *et al.*, 2023). NFT's are digital assets on the blockchain that function like cryptocurrencies (Lee *et al.*, 2023) Immutable X enables gas-free NFT trading, eliminating the need for users to pay transaction fees (gas) (Huynh-The *et al.*, 2023; Smith *et al.*, 2023).

Decentraland utilises MANA tokens, and Sandbox uses SAND tokens, both on the Ethereum blockchain. MANA is used for virtual land creation and trading, while SAND is employed for creating, owning, and monetizing gaming experiences (Huynh-The *et al.*, 2023).

Axie Infinity is a play-to-earn metaverse project, enabling players to boost their digital assets by advancing in-game abilities to specific levels. Axie Infinity has two tokens, Axie Infinity Shards (AXS) and Small Love Potion (SLP) tokens. Players can claim AXS rewards by staking their tokens, participating in governance, and creating user-generated content. SLP is granted through in-game activities. Most transactions occur on the Ronin sidechain, connected to Ethereum, to reduce fees (Huynh-The *et al.*, 2023).

2.2 Existing Reward Systems

Based on the methodology employed, this section categorizes the literature into two parts: reactions to rewards and reward types.

2.2.1 Reaction to rewards

Social comparison theory, developed by psychologist Leon Festinger in 1954, explains how individuals gauge their abilities, opinions, and emotions by comparing themselves to others. This comparison can be upward, where individuals compare themselves to those they perceive as better to improve themselves, or downward, where they compare to those worse off to feel better about their abilities. Social comparison leads to competitive behaviour and higher motivation (Christy and Fox, 2014; Dong and Zhu, 2023). Loss aversion is the tendency to avoid losses over equivalent gains but requires that the potential loss has value to the individual, such as monetary or psychological worth earned through effort (Lin *et al.*, 2023).

2.2.2 Reward types

Pombo et al. (2023) identified reward systems using points, leaderboards, and badges to recognise performance and progress. Points are granted for tasks or skills, leaderboards rank participants, and badges symbolise task completion (Pombo and Santos, 2023). Point-based incentives enhance employee motivation (Farzan *et al.*, 2008; Pombo and Santos, 2023), and leaderboards improve motivation and approval rates (Pombo and Santos, 2023). Leaderboards stimulate competition and social comparisons in a business context (Christy and Fox, 2014).

Badge-based systems boost engagement and commitment to task completion when aligned with objectives, promoting intrinsic motivation (Pombo *et al.*, 2019; Pombo and Santos, 2023). Reward systems can encompass both tangible and intangible elements, combining awards with a sense of achievement (Pombo and Santos, 2023).

2.3 Blockchain Technology

In this section, a literature review in blockchain technology is conducted, to identify techniques ensuring confidentiality, integrity, availability, and security risks.

Blockchains are decentralised databases that store a registry of transactions and are primarily utilised in cryptocurrency platforms but have applications in diverse sectors such as Healthcare, Real Estate, Supply Chain, Internet-of-Things (IoT), Government, Cyber Security, Social Media and Artificial Intelligence (AI) (Zantalis *et al.*, 2022; Raj, 2023). Their decentralised nature eliminates the need for a trusted third party (Li *et al.*, 2020; Zantalis *et al.*, 2022). Blockchain technology makes use of cryptography, Merkle tree, difficulty level and consensus algorithms, which are explained below.

2.3.1 Technology utilised

- Cryptography: Used for secure communication and data immutability in blockchain. It transforms data into ciphertext through encryption, which can be reversed with a key. Blockchain uses symmetric-key cryptography (one key) and asymmetric key cryptography (a key pair) to achieve this security (Mallouli *et al.*, 2021).
- Merkle tree: A hash binary tree for rapid integrity verification of data (Zhai *et al.*, 2019), confirming transaction presence in a block without requiring the entire block or transaction data (Samuel, 2015; Zantalis *et al.*, 2022).
- Difficulty level: Regulates the block creation rate, particularly in Proof-of-Work (PoW) blockchains, where miners aim to find a hash below a set threshold. PoW is a type of consensus algorithm and is discussed next. Higher difficulty levels deter attacks like the 51% attack, where one entity controls over 50% of network power, enabling manipulation and disruptions. Difficulty level ensures consistent block creation, balances mining power, incentivises honest participation, and maintains network stability (Mardiansyah and Sari, 2022; Academy, 2023).
- Consensus algorithms: Trust in blockchain transactions relies on mathematical algorithms, collectively known as the consensus mechanism (Samuel, 2015). Blockchain relies on decentralised consensus mechanisms for reliability and consistency (Li *et al.*, 2020). They address the Byzantine Generals' Problem (BGP) in distributed computing, which requires trust even in the presence of malicious nodes. Byzantine Fault Tolerant (BFT) algorithms are designed to handle trust challenges in distributed computing and achieve consensus, even in the presence of adversarial nodes (Lamport *et al.*, 2019; Zhang *et al.*, 2019). Consensus is possible if fewer than 1/3 of nodes are faulty by way of majority voting (Samuel, 2015). However, BFT has high communication complexity and relies on the majority node honesty (Zhang *et al.*, 2019). Popular consensus algorithms identified from research, detailed in Table 1, can resist BGP, provided the 51% attack doesn't happen.

2.3.2 Known attacks to blockchains

Some known attacks to blockchains include:

- 51% Attack: As mentioned, when someone controls over 50% of a blockchain's computational power to manipulate transactions. Higher difficulty levels make it more challenging and costly, potentially serving as a mitigation.
- Sybil Attack: Seizing control of multiple network identities to exert undue influence. PoW serves as defence (Lee *et al.*, 2022).
- Distributed Denial-of-Service (DDoS) Attack: Strain the network with requests to render it unavailable.
- Selfish-mining Attack: Withholding discovered blocks to waste other miners' computational power. The attacker accumulates 2N blocks and releases them when other miners find more than K (where N > K) blocks, disrupting the main chain. Countermeasures focus on block publication delay (Lee *et al.*, 2022).
- Nothing-at-Stake Attack: Occurs during forks when an attacker adds their block to multiple branches to boost its chance of becoming the next valid block. Forks occur when multiple blocks are added to the previous block. In PoW, shared resources decrease a miner's chance of finding a new block,

making this attack unlikely. Mitigation includes introducing penalties in PoS algorithms (Ferdous *et al.*, 2021).

- Bribing (Short-range) Attack: An attacker attempts to double-spend by creating a fork, e.g., an attacker buys goods, waits for delivery after a specific number of blocks, then forks the chain and bribes miners to make the forked branch longer. This reverses the transaction, enabling the attacker to spend the currency again. If the bribe cost is less than the goods price, it is profitable for the attacker. Mitigation involves penalty mechanisms (Ferdous *et al.*, 2021).
- Long-range Attack: An attacker constructs an alternative blockchain, colluding with stakeholders and seeking double-spending opportunities. Check-pointing, ensuring the chain becomes irreversible up to a certain block, serves as mitigation (Ferdous *et al.*, 2021).
- Coin-age Accumulation Attack: Manipulating algorithms with accumulated coin-age. Coin-age is trust in the authenticity of a miner by the period his coins have been kept as stake. A cap on coin-age serves as defence (Ferdous *et al.*, 2021).
- Cartel-formation Attack: A small group dominating mining pools, potentially launching 51% attacks in PoW or exploiting stakes in PoS algorithms (Ferdous *et al.*, 2021).
- Stake-grinding Attack: Allows an attacker to increase the probability of generating subsequent blocks if there's insufficient randomness in PoS algorithms. Mitigation involves establishing a secure source of randomness (Ferdous *et al.*, 2021).

Table 1: Consensus algorithms (Samuel, 2015; Zhang *et al.*, 2019; Manolache *et al.*, 2021)

Consensus Algorithms	Limitations	Properties
Proof-of-Work (PoW)	1. High computation complexity, low success probability. 2. Energy consumption in hash computation. 3. Mining centralization to boost block mining chances and risk of 51% attack. 4. Tragedy of commons: miners leaving due to low rewards.	1. Computationally difficult yet easily verifiable puzzles for data credibility verification.
Proof-of-Stake (PoS)	1. 51% attack possible when one user controls over 50% of the blockchain. 2. No mining rewards. 3. Transaction fees divided among validators, potentially discouraging block mining with low fees.	1. Ensures CAP Theorem. 2. Reduces dependence on rewards with penalty-based solutions. 3. Requires miners to stake their capital to qualify for block mining.
Delegated Proof-of-Stake (DPos)	1. 51% attack risk with insufficient validators. 2. Coin-wealth reliance leads to disproportionate influence.	1. Reputation and user voting select validators.
Proof-of-Authority (PoA)	1. Vulnerable to DDoS attacks (described in section below). 2. Relies on validators' identity verification. 3. Requires investments in money and reputation to mitigate risks. 4. 51% attacks are more challenging.	1. A cost-effective PoS variant with identifiable block creators. 2. Predictable block production through "target spacing", where blocks are produced on fixed time intervals.

Most attacks occur during block mining competitions and result from vulnerabilities in the consensus algorithms being exploited. It's worth noting that PoS variants are more susceptible to attacks, but implementing PoW requires more computational power to maintain a sufficiently high difficulty level. In the proposed model, an attempt is made to address this weakness.

The next section proposes a metaverse reward model.

3. Proposed Metaverse Reward Model

In this section, a metaverse reward model is presented, building on the findings in the previous section. It includes a conceptual model, individual component flow diagrams, and a comprehensive model flow diagram.

3.1 High-Level View of Proposed Metaward Model

Each of the components in the conceptual model are elaborated below, as represented in Figure 1:

- Tasks or challenges are generated and reviewed with rule implementation, ensuring alignment with the organization's requirements.
- Once approved, tasks are assigned to relevant employees with specified deadlines and notifications. Periodic reminders are sent to motivate individuals who are driven by loss aversion to complete their tasks promptly and avoid penalties.
- Failure to complete tasks results in a cryptocurrency penalty deducted from the employee's wallet.
- Employees receive cryptocurrency rewards upon timely and successful task completion, accumulating in their wallet. Rewards have expiry dates for regular redemption. The organization offers flexibility by allowing various redemption options, such as converting cryptocurrency to fiat currency, receiving instant rewards, taking time off, and more. Employees can accumulate rewards until they have a sufficient balance for their preferred redemption.
- To foster social comparison, employees receive icon-like badges upon completing tasks. Similar to rewards, badges have expiration dates to ensure accurate and up-to-date recognition of achievements.
- Employee rankings fluctuate based on completed/non-completed tasks, appealing to those motivated by social comparison. An expiry date ensures fair ranking changes with time and task completion.

The figure below represents the conceptual model:

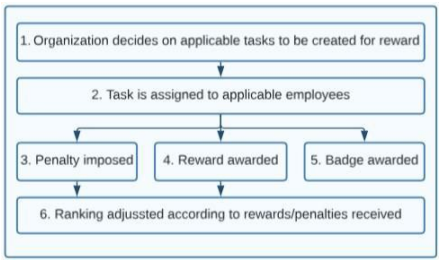


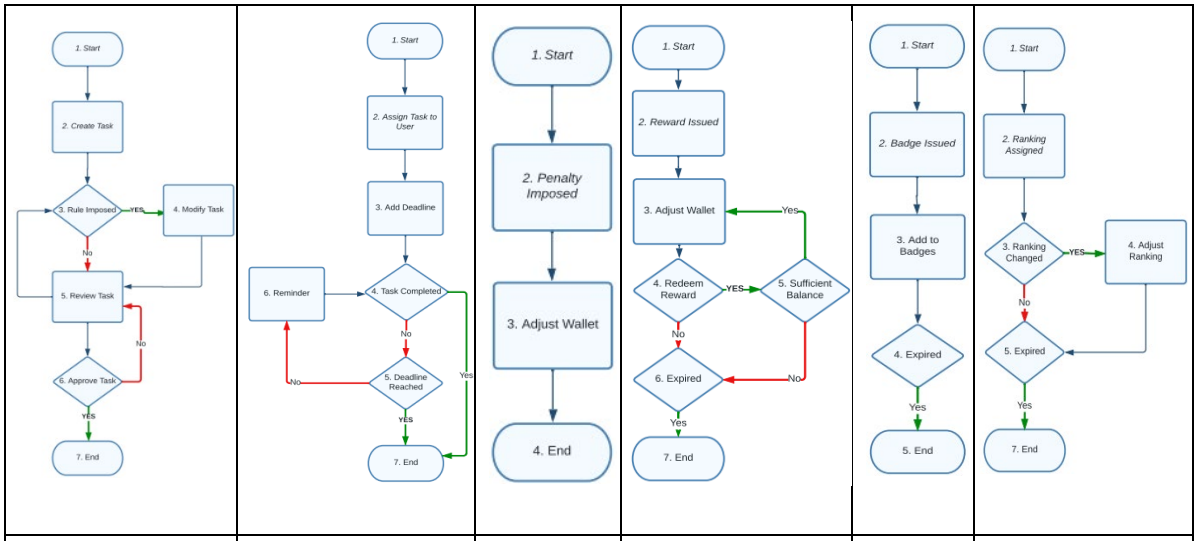
Figure 1: Conceptual Metaward model

3.2 Detailed View of the Metaward Model

This section illustrates each component of the Metaward Model, describing events within each. Figure 2 depicts the component diagrams.

Task created (Figure 2.1):

- Start - The beginning of Metaward model, either after login or when integrated into an existing setup.
- Create Task - Tasks or challenges are generated in this step.
- Rule Imposed - This checkpoint handles rule implementation. It proceeds to "Modify Task" for further rule implementation or "Review Task" if no additional changes are needed.
- Modify Task - Task modifications are made in this phase.
- Review Task - After implementing modifications and rules, the task undergoes review.
- Approve Task - If approved, it moves to the next stage; otherwise, it returns to "Review Task" for further assessment and modifications.
- End – Placeholder to mark the end of the "Task Created" process. In the complete model, the task or challenge will be assigned from this point.



1 Task created	2 Task assigned	3 Penalty	4 Reward	5 Badge	6 Ranking

Figure 2: Component Diagrams

Task assigned (Figure 2.2):

- Start - Placeholder to begin the "Task Assigned" component, to be removed in the complete model.
- Assign Task - The task is allocated to an employee or user.
- Add Deadline - A completion deadline is set during this step.
- Task Completed - A checkpoint for periodic checks to verify task or challenge completion. If not completed, it proceeds to a checkpoint to confirm if the deadline has passed.
- Deadline Reached - A checkpoint to confirm if the deadline has been exceeded.
- Reminder - If the deadline has not passed, a reminder is sent to prompt timely task completion.
- End - Placeholder to mark the end of the "Task Assigned" process. In the complete model, it will move to either the "Reward" or "Penalty" process, depending on task completion status.

Penalty (Figure 2.3):

- Start - Placeholder to commence the "Penalty" component, to be removed in the complete model.
- Penalty Imposed - Imposition of a penalty for failure to complete the task timeously.
- Adjust Wallet - The user's wallet is deducted with a negative cryptocurrency amount for the penalty.
- End - Placeholder to mark the end of the "Penalty" process, to be removed in the complete model.

Reward (Figure 2.4):

- Start - Placeholder to commence the "Reward" component, to be removed in the complete model.
- Reward Issued - Issuance of a reward in the form of cryptocurrency.
- Adjust Wallet - Modification of the employee's balance to reflect the added reward.
- Redeem Reward - Checkpoint for potential reward redemption.
- Sufficient Balance - Verification of adequate funds. If balance is sufficient, the reward is redeemed, and the wallet is adjusted.
- Expired - Handling of expired rewards.
- End - Denotes the conclusion of the process.

Badge (Figure 2.5):

- Start - Placeholder to begin the "Badge" component, to be removed in the complete model.
- Badge Issued - Issuance of a badge for completed tasks.
- Add to Badges - Inclusion of the badge in the employee's collection.
- Expired - Removal of expired badges from the collection.
- End - Marks the conclusion of the process.

Ranking (Figure 2.6):

- Start - Placeholder to commence the "Ranking" component, to be excluded in the complete model.
- Ranking Assigned - Allocation of a ranking based on task completion.
- Ranking Changed - Verification of any changes in the employee's leaderboard position.
- Adjust Ranking - If the employee's ranking has changed, the leaderboard is updated accordingly.
- Expired - Removal of expired rankings.
- End - Marks the conclusion of the process.

3.3 Comprehensive Metaward System

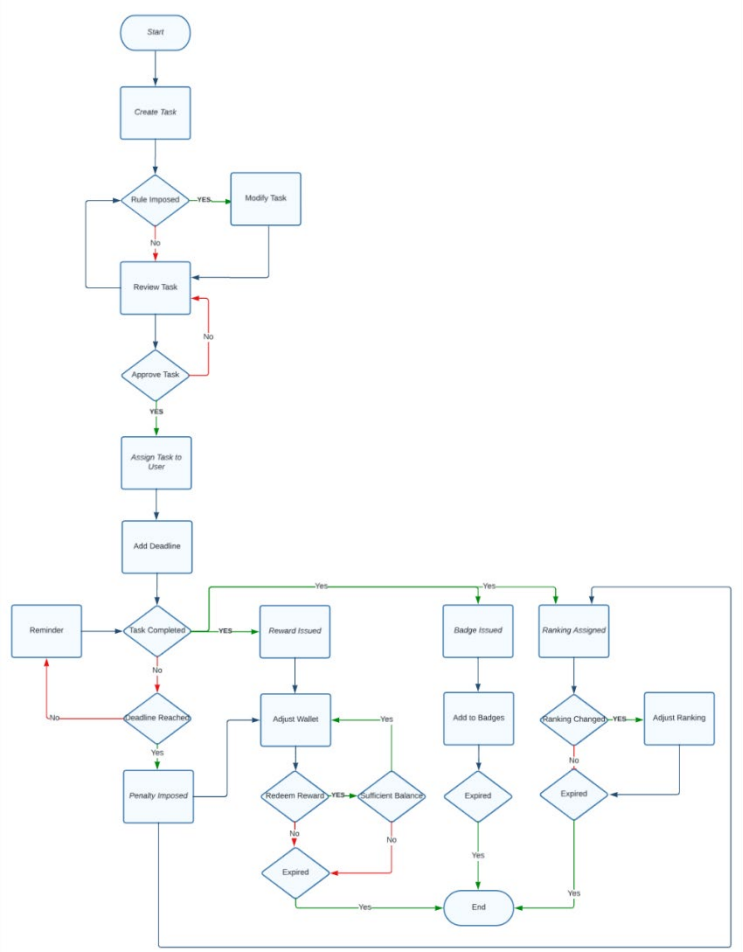


Figure 3: Metaward reward system

Figure 3 above, depicts the inclusive Metaward system, addressing various employee motivations with badges for recognition-seeking employees, leaderboards for those motivated by social comparison, rewards for those driven by incentives, and timed penalties caters to individuals preferring loss aversion. Initial rewards are in cryptocurrency, offering flexibility for later redemption in line with the organization's choices, including monetary rewards, instant benefits, or time off.

The following section presents the proof of concept.

4. Proof of Concept

In the proof of concept, two blockchain solutions are introduced, one using PoW and the other employing the (MultiversX, 2018) Secure Proof-of-Stake (SPoS), a PoS variant. Both randomise block creation to eliminate mining competition. The PoW mechanism offers robust security but demands significant computational power to maintain a high difficulty level. However, it lacks flexibility when changing tasks or challenges. On the contrary, the SPoS system, incorporates an innovative solution by MultiversX (2018b), that simultaneously addresses state, network, and transaction sharding. With its adaptive component, this architecture allows dynamic

network configuration, ensuring both high security and scalability to meet demand. Additionally, new challenges or tasks can be deployed through smart contracts on the network.

To evaluate the system, a small-scale test with five hypothetical employees is conducted. Each receives an initial 100 Metagems (the Metaward's cryptocurrency), in their wallets. These Metagems can only be redeemed once the employee has earned the equivalent amount by completing tasks. However, non-completion results in Metagems being deducted from their wallet and still requires the employee to earn 100 Metagems before redeeming a reward. Additionally, negative balances are allowed in the wallet. An expiry date of six months is applied to rewards and badges, and employees can earn multiple badges of the same type. A penalty higher than the reward aims to encourage employees to prioritise avoiding loss over earning rewards. The reward structure is designed to be easily attainable, allowing for the accumulation and redemption of Metagems before expiration. The leaderboard highlights top-performing employees, fostering a culture of excellence.

Various tasks are deployed, assuming that the organisation is a remote tech company: 1) Employees receive 50 Metagems and an "Innovator" badge for suggesting new product features or solving technical challenges, with no penalties; 2) Employees receive 20 Metagems and a "Deadline Dynamo" badge for timely project completion but face a penalty of 30 Metagems for untimely completion; 3) Employees earn 10 Metagems and a "Virtual Virtuoso" badge for attending virtual meetings, with a 15 Metagems penalty for missed meetings; 4) Leaderboards are displayed upon employee login to foster healthy competition among development teams.

Reward redemption options are set as follows: 1) 200 Metagems – a R200 lunch voucher at the employee's choice restaurant; 2) 300 Metagems – movie tickets for two; 3) 500 Metagems – a day off for the employee.

Test Cases: 1) Employee 1 proposed an innovative solution, attended five virtual meetings, and completed the project on time, earning 120 Metagems, an "Innovator" badge, a "Deadline Dynamo" badge, and five "Virtual Virtuoso" badges. This allowed redemption of a lunch voucher, leaving a wallet balance of 20 Metagems after the reward; 2) Employee 2 attended five virtual meetings, earning 50 Metagems and five "Virtual Virtuoso" badges, resulting in a wallet balance of 150 Metagems, which is insufficient for reward redemption; 3) Employee 3 proposed an innovative solution and solved a complex technical issue, earning 100 Metagems and two "Innovator" badges. Employee 3 chose not to redeem rewards, maintaining a wallet balance of 200 Metagems; 4) Employee 4 attended five virtual meetings but missed the project deadline, earning 50 Metagems and five "Virtual Virtuoso" badges. A penalty of 30 Metagems was imposed for untimely completion, resulting in a wallet balance of 120 Metagems, insufficient for reward redemption; 5) Employee 5 missed three meetings, incurring a penalty of 45 Metagems, leaving a wallet balance of 55 Metagems. Since Employee 5 has not earned Metagems, they cannot redeem any rewards and have lost 45 of their initially awarded Metagems.

The rankings after the experiment are as follows: 1) Employee 1; 2) Employee 3; 3) Employee 2; 4) Employee 4; 5) Employee 5.

The following section presents the related literature.

5. Related Literature

This research's primary objective is to find an effective reward system for motivating and retaining employees in the Metaverse. This section reviews relevant literature on reward systems to inform this goal.

Notably, the study found only one enterprise-level reward system, the focus of this research. Farzan et al. (2008) implemented a points-based incentive system in a social networking site for IBM employees. It effectively motivated user contributions, particularly when combined with email notifications. Visible points encouraged more content, discussions, and allowed site owners to influence user behaviour for specific goals, benefiting marketing and content generation. However, some users wanted more control over the visibility of their activity due to privacy concerns. Christy and Fox, (2014) implemented a leaderboard reward system in educational settings, which impact academic performance and identity. However, the study contained a limited sample, requiring further research. Pombo et al. (2019) and Pombo and Santos (2023) introduced badge-based systems in educational contexts. In the study conducted by Pombo et al. (2019) at two Portuguese universities, gamification enhanced education by increasing engagement, motivation, and alignment with learning goals. Pombo and Santos (2023) used a badge-based system for in-class activities, following SCRUM principles over five sprints, featuring student achievements, class creation, and badge rewards. The findings can guide researchers and practitioners in implementing similar tools for in-class activities. Lin et al. (2023) employed a loss aversion strategy in an outdoor AR exergame, leading to longer, more intense exercise sessions compared to a reward-based approach. However, it caused pressure for one player. Study suggests careful use of loss aversion to avoid

discouraging players. Future research should factor in personality traits, develop adaptive games with better content, usability, multiplayer options, and adjust exercise intensity based on fitness levels.

Each of these studies applied a unique reward strategy and analysed its individual outcomes, in contrast to the Metaward system, which integrates various motivational and incentivising elements. The next section will delve into the Metaward reward system.

6. Discussion

The Metaward reward system offers a comprehensive solution to enhance employee motivation and engagement. It leverages a combination of rewards, penalties, recognition, and social comparison to create a dynamic and motivating work environment. Organisations can customise this approach to align with their goals and values, ultimately contributing to improved employee performance and satisfaction.

Badges are awarded to employees upon task completion, serving as a form of recognition. They are used to acknowledge achievements and motivate those seeking recognition. Badges also expire for accuracy and up-to-date recognition. Employee rankings fluctuate based on task completion, catering to individuals motivated by social comparison. These rankings are periodically adjusted to ensure fairness and accuracy over time. Rewards are provided to employees upon successful and timely task completion. These rewards accumulate in their wallets and can be redeemed for various options, such as fiat currency, instant benefits, or time off. This flexibility allows employees to choose rewards suited to their preferences. The rewards also expire to ensure regular redemption and sustainability in issuing rewards from the organisational perspective. In addition to rewards, the penalty component introduces a loss aversion element. If employees fail to complete tasks timeously, they face cryptocurrency penalties deducted from their wallets. Periodic reminders help motivate individuals to avoid penalties by completing tasks promptly.

In the subsequent section, this research concludes its findings and outlines potential avenues for future work.

7. Conclusion

The Metaward reward system presents a comprehensive framework to invigorate employee motivation and engagement. By amalgamating rewards, penalties, recognition, and social comparison, it constructs a dynamic work environment that can be tailored to harmonise with an organisation's mission and values. This multifaceted approach has the potential to significantly enhance employee performance and job satisfaction.

Several avenues for future work warrant exploration in: 1) Long-term impact on motivation and performance.; 2) Adaptability across different industries; 3) Use of gamification to boost user engagement; 4) Continuous improvement based on employee feedback; 5) Suitability in diverse cultural contexts; 6) Addressing ethical concerns like privacy and fairness. These future directions hold the potential to further enhance the Metaward system's effectiveness and expand its applicability as a powerful tool for bolstering employee motivation, engagement, and overall workplace satisfaction.

References

Abdullah, N.A.A. et al. (2020) 'COPING WITH POST COVID-19: CAN WORK FROM HOME BE A NEW NORM?' European Journal of Social Sciences Studies, 5(6). DOI: 10.46827/ejsss.v5i6.933.

Academy, B. (2023) Difficulty | Binance Academy. Available at: https://academy.binance.com/en/glossary/difficulty

Ali, B.J. and Anwar, G. (2021) 'An Empirical Study of Employees' Motivation and Its Influence Job Satisfaction'. DOI: 10.22161/ijebm.5.2.

Christy, K.R. and Fox, J. (2014) 'Leaderboards in a Virtual Classroom: A Test of Stereotype Threat and Social Comparison Explanations for Women's Math Performance'. Computers and Education, 78, pp. 66–77. DOI: 10.1016/j.compedu.2014.05.005.

Dong, Y. and Zhu, Q. (2023) 'How Downward Social Comparison Motivates Workers: A Structural Equation Model on Personality, Social Comparison Orientation, Motivation and Performance'. In ACM International Conference Proceeding Series. Association for Computing Machinery, pp. 207–212. DOI: 10.1145/3587889.3588201.

Farzan, R. et al. (2008) When the Experiment Is over: Deploying an Incentive System to All the Users.

Ferdous, M.S., Chowdhury, M.J.M. and Hoque, M.A. (2021) DOI: 10.1016/j.jnca.2021.103035.

Hussain, S.D. et al. (2019) 'Impact of Employees' Recognition, Rewards and Job Stress on Job Performance'. SEISENSE Journal of Management, 2(2), pp. 69–82. DOI: 10.33215/sjom.v2i2.121.

Huynh-The, T. et al. (2023) 'Blockchain for the Metaverse: A Review'. Future Generation Computer Systems, 143, pp. 401–419. DOI: 10.1016/j.future.2023.02.008.

Keshav, S. and Cheriton, D.R. (2007) How to Read a Paper. Available at: http://www.sigcomm.org/ccr/drupal/

Lamport, L., Shostak, R. and Pease, M. (2019) 'The Byzantine Generals Problem'. In Concurrency: The Works of Leslie Lamport. New York, NY, USA: Association for Computing Machinery, pp. 203–226. Available at: https://doi.org/10.1145/3335772.3335936

Lee, C.T., Ho, T.Y. and Xie, H.H. (2023) 'Building Brand Engagement in Metaverse Commerce: The Role of Branded Non-Fungible Toekns (BNFTs)'. Electronic Commerce Research and Applications, 58. DOI: 10.1016/j.elerap.2023.101248.

Lee, S., Lee, D. and Kim, S. (2022) 'Block Double-Submission Attack: Block Withholding Can Be Self-Destructive'. In Proceedings of the 4th ACM Conference on Advances in Financial Technologies. New York, NY, USA: ACM, pp. 232–243. DOI: 10.1145/3558535.3559787.

Li, X. et al. (2020) 'A Survey on the Security of Blockchain Systems'. Future Generation Computer Systems, 107, pp. 841–853. DOI: 10.1016/j.future.2017.08.020.

Lin, Y. et al. (2023) 'Dragon Hunter: Loss Aversion for Increasing Physical Activity in AR Exergames'. In Association for Computing Machinery (ACM), pp. 212–221. DOI: 10.1145/3579375.3579403.

Mallouli, F. et al. (2021) 'Cryptocurrency Bounced Back Based on Cryptography Technology during the COVID-19 Pandemic'. In Proceedings - 2021 International Conference on Computational Science and Computational Intelligence, CSCI 2021. Institute of Electrical and Electronics Engineers Inc., pp. 609–614. DOI: 10.1109/CSCI54926.2021.00031.

Manolache, M.A., Manolache, S. and Tapus, N. (2021) 'Decision Making Using the Blockchain Proof of Authority Consensus'. In Procedia Computer Science. Elsevier B.V., pp. 580–588. DOI: 10.1016/j.procs.2022.01.071.

Mardiansyah, V. and Sari, R.F. (2022) 'SimBlock Simulator Enhancement with Difficulty Level Algorithm Based on Proof-of-Work Consensus for Lightweight Blockchain'. Sensors, 22(23). DOI: 10.3390/s22239057.

MultiversX. (2018) Welcome to MultiversX • MultiversX Docs. Available at: https://docs.multiversx.com/welcome/welcome-to-multiversx/

Pombo, N., Garcia, N. and Alves, P. (2019) How to Get a Badge? Unlock Your Mind. ieee educon. IEEE.

Pombo, N. and Santos, H. (2023) 'Lessons Learned from the Development of a Computerised Badge-Based Reward Tool for Student Engagement in Learning Activities'. In EDUNINE 2023 - 7th IEEE World Engineering Education Conference: Reimaging Engineering - Toward the Next Generation of Engineering Education, Merging Technologies in a Connected World, Proceedings. Institute of Electrical and Electronics Engineers Inc. DOI: 10.1109/EDUNINE57531.2023.10102849.

Raj, R. (2023) 10 Real-World Blockchain Applications | Intellipaat. Available at: https://intellipaat.com/blog/tutorial/blockchain-tutorial/blockchain-applications

Roose, K. (2021) Why Did Facebook Become Meta? - The New York Times. DOI: https://www.nytimes.com/2021/10/29/technology/meta-facebook-zuckerberg.html

Samuel, S. (2015) Blockchain Technology: A Quick Introduction | UpGrad Free Course. Available at: https://www.upgrad.com/za/introduction-to-blockchain-technology-free-course/

Scott, G. (2022) Linden Dollar: What It Is, How It Works, Tax Implications. Available at: https://www.investopedia.com/terms/l/linden-dollar.asp

Smith, C. et al. (2023) Gas and Fees | Ethereum.Org. Available at: https://ethereum.org/en/developers/docs/gas/

Xu, M. et al. (2023) 'A Trustless Architecture of Blockchain-Enabled Metaverse R'. 3, p. 88. DOI: 10.1016/j.hcc.2022.10.

Young, W. et al. (2015) 'Changing Behaviour: Successful Environmental Programmes in the Workplace'. Business Strategy and the Environment, 24(8), pp. 689–703. DOI: 10.1002/BSE.1836.

Zantalis, F., Koulouras, G. and Karabetsos, S. (2022) 'Blockchain Technology as a Framework Blockchain Technology: A Framework for Endless Applications'. DOI: 10.1109/MCE.2022.Doi.

Zhai, S. et al. (2019) 'Research on the Application of Cryptography on the Blockchain'. In Journal of Physics: Conference Series. Institute of Physics Publishing. DOI: 10.1088/1742-6596/1168/3/032077.

Zhang, R., Xue, R. and Liu, L. (2019) 'Security and Privacy on Blockchain'. ACM Computing Surveys, 52(3). DOI: 10.1145/3316481.

Cybersafe: Gamifying Cybersecurity Training with a Training App

Carlos Roque[1], Gareth Moodley[2] and Sayonnha Mandal[2]
[1]School of Engineering, UAGM Gurabo Campus, Puerto Rico
[2]College of Information Science and Technology, University of Nebraska Omaha, USA

croque16@email.uagm.edu
gmoodley@unomaha.edu
smandal@unomaha.edu

Abstract: The rapidly evolving digital landscape has triggered a surge in cybersecurity threats, particularly social engineering techniques which demand innovative and accessible countermeasures to combat them in an accessible and real-time format. We present "Cybersafe," a mobile application designed to empower users to identify and combat common social engineering exploits effectively. The transformative concept behind this initiative aims to reshape traditional cybersecurity training by introducing a gamified, user-friendly platform suitable for all age groups. Cybersafe's application's functionality revolves around interactive quizzes that assess users' ability to identify threats. The findings from the research serve as a valuable resource for cybersecurity trainers, application developers, and organizations striving for a secure digital environment.

Keywords: Cyber security, Mobile devices, Phishing, Social engineering,, Gamify, Vulnerability

1. Introduction

The rapidly evolving digital landscape has triggered a surge in cybersecurity threats, particularly social engineering techniques which demand innovative and accessible countermeasures to combat them in an accessible and real-time format. Current similar applications, such as NightHawk (Liu, et. al 2022) by PhishFort and GoPhish (Luse et. al, 2021), aim to actively prevent phishing by way of anti-phishing software and by assessing the security posture of enterprise employees trying to expose them to social engineering attacks. However, they both lack in teaching the user how to prevent phishing. They do not offer a set of instructions on how to observe behaviour from social inquiries, which can effectively plant ideas for end-users on how to catch a phishing attempt quickly and responsibly. Our app features a questionnaire-based approach categorized into different social engineering threats, where each question is assigned a specific weight. The user responds to the questions depending on the specific operational context and a cumulative score is assessed through the application to report the final threat likelihood score. Moreover, the app provides specific recommendations in a user-friendly format which enables the user to implement actionable items to change their practices. By adopting a gamified approach, Cybersafe intends to overcome the challenges of user engagement and retention typically associated with conventional training methods. The development and deployment of Cybersafe signifies a meaningful shift in cybersecurity training methods. Each question within the quiz holds a specific score and answering "yes" to a question indicates a potential vulnerability, subsequently increasing the total threat likelihood score. Broad questions are asked first, followed by more focused ones. Then, recommendations are shown that give the user actionable items to change their practices.

This scoring system serves to guide users in distinguishing genuine emails from phishing attempts, enhancing their cybersecurity skills progressively. The app also extends its learning scope to include web security, covering a broad spectrum of cybersecurity aspects to ensure comprehensive user awareness.

The paper further explores the developmental aspects of the Cybersafe application and provides an analysis of user feedback and learning outcomes. It concludes with a summary of findings and the potential impact of such a tool on global phishing prevention efforts.

2. Background

2.1 Business Email Compromise Attacks

The rising severity of phishing and other cyber threats is well documented. A 2019 report from the FBI highlighted the scale of the issue, suggesting that Business Email Compromise (BEC) attacks alone resulted in estimated damages of around 26 billion dollars. These attacks typically involve the impersonation of a business executive or an organization with the aim of tricking employees, partners, or customers into transferring money or sensitive data (Bakarich et. al., 2019).

2.2 COVID-Themed Phishing Lures

The threat has only escalated with recent global events such as the COVID-19 pandemic, which saw an increase in COVID-themed phishing lures. Cybercriminals quickly took advantage of the uncertainty and fear surrounding the pandemic, crafting deceptive emails and messages to trick users into revealing personal information or downloading malicious files.

2.3 Time Pressure Techniques

In addition, cybercriminals often employ "time pressure" psychological techniques in their attacks. These techniques create a sense of urgency that can cause users to react without taking the time to verify the legitimacy of the request. This further underscores the need for users to be educated about such tactics and the importance of verifying information before acting.

2.4 Risk-Taking

Risk-taking behaviour also plays a significant role in users' vulnerability to cyber threats. Users who are unaware of the risks associated with certain online behaviours, or who underestimate these risks, are more likely to fall victim to cyber-attacks (Aleroud et. al., 2017) .

The increasing sophistication and scale of these threats highlight the importance of tools like Cybersafe. By educating users on the risks and signs of phishing and other cyber threats, Cybersafe aims to reduce the likelihood of users falling victim to these attacks.

3. Adversary Motivations

Understanding the motivations of cyber attackers can provide valuable insights into the types of threats users may face and the tactics attackers may use. Common motivations include:

3.1 Financial Gain

One of the primary motivations for cyber-attacks is financial gain. Cybercriminals often target personal and financial information that can be used for fraudulent purposes or sold on the black market. Attacks like phishing and ransomware are typically financially motivated, aiming to trick users into providing sensitive information or paying a ransom (Abroshan et. al., 2021).

3.2 Information Stealing and Fraud

Cyber attackers may also be motivated by the desire to steal information. This can include personal data, business information, intellectual property, and even state secrets. The stolen information can be used for a range of malicious activities, from identity theft to corporate espionage (Chaudhary, 2016).

Fraud is another common motivation. Cybercriminals may impersonate legitimate entities to trick users into providing sensitive information or manipulate systems and transactions for illicit gain.

3.3 Spying Campaigns

Some cyber-attacks are part of broader spying campaigns, often conducted by state-sponsored actors. These attacks aim to gather intelligence on individuals, organizations, or nations for political or strategic purposes.

3.4 Political Motivations

Political motivations can also drive cyber-attacks. These can range from hacktivist campaigns aimed at promoting a political cause, to state-sponsored cyber warfare aimed at destabilizing or discrediting an adversary.

By understanding these motivations, Cybersafe can better equip users to recognize and respond to potential threats, reducing their risk of falling victim to cyber-attacks.

4. Related Research

Anti-phishing applications have been proposed and utilized since the beginning of phishing campaigns. Fairly recent research includes apps focused solely on email phishing (Helmi et. al., 2019), phishing using Wi-Fi hotspots (Chen et. al., 2021)) and specifically phishing on handheld mobile devices (Chorge et. al., 2016). Several efforts have been made to incorporate user awareness (Zieni et. al., 2023) components in phishing apps as well as apps dedicated primarily to educate users (Dixon et. al., 2019) regarding the various phishing and web exploits via different mediums (Lungu et. al., 2010). Applications such as MobiFish (Wu et. al., 2014) and NoPhish (Canova

et. al., 2014) have been proposed in prior research. For discussion in this section, we have identified at least two other similar candidates that attempt to prevent social engineering attacks and actively teach the user, these are: NightHawk by PhishFort, and GoPhish.

NightHawk, developed by PhishFort, is a robust platform specifically designed to combat phishing threats. It uses machine learning to automatically detect, analyse, and respond to phishing attacks in real time (Liu et. al., 2022). Users receive immediate alerts about potential threats and get comprehensive reports about each attack to deepen their understanding and improve future defence strategies. The operations for analysing URLs and detecting potential phishing links are performed at the backend of the app. The users are typically presented with a user interface with a custom watermark to denote that the URL is safe to proceed. The application lacks the ability to take the users step by step through the process of reasoning behind the final decision. The user only views the end user interface and the final recommendation. Hence, users remain unaware of potential phishing markers that are needed to identify similar future vulnerabilities. Moreover, NightHawk focuses more on the detection and immediate response part. Its user interface is designed for security teams rather than average internet users, which makes it less accessible for the public. On the other hand, Cybersafe emphasizes the educational element for the user to better understand social engineering attacks and potentially identify similar future attacks.

GoPhish is an open-source phishing toolkit designed for businesses and security professionals to test and enhance their organization's defences against phishing (Luse et. al., 2021). It allows users to create simulated phishing attacks to identify vulnerabilities within the organization. The tool then provides detailed reports about the simulation results to help educate the users and develop more effective strategies for real phishing attacks. GoPhish adopts a higher-level approach to scan and detect all potential phishing vulnerabilities in the users' systems. It provides a comprehensive analysis of user operations, such as, clicked URLs, data submitted, emails viewed, to provide the user with a view of their online activity and resultant exploits that they fell prey to. GoPhish's primary target audience is organizations rather than individual users, which contrasts with Cybersafe's emphasis on individual internet users. While both Cybersafe and GoPhish share the common goal of educating users about phishing attacks, their approaches differ significantly. GoPhish is a proactive tool aimed at training users through simulation, while Cybersafe is more reactive, focusing on real-time detection and education of users as threats occur.

In conclusion, while NightHawk and GoPhish bear similarities to Cybersafe, they differ in key aspects such as their target audience, usage approach, and primary focus. This makes Cybersafe a unique tool in its approach to preventing social engineering attacks and educating users.

5. Goals of Cybersafe

In an increasingly digital world, phishing attacks have become all too common, claiming many victims who lack the necessary knowledge to identify and avoid such threats. Users often inadvertently share sensitive information, such as login credentials or personal details, without recognizing the signs of a phishing attempt (Figure 1). This vulnerability is largely due to insufficient awareness or understanding of how phishing operates and the potential signs of an attack.

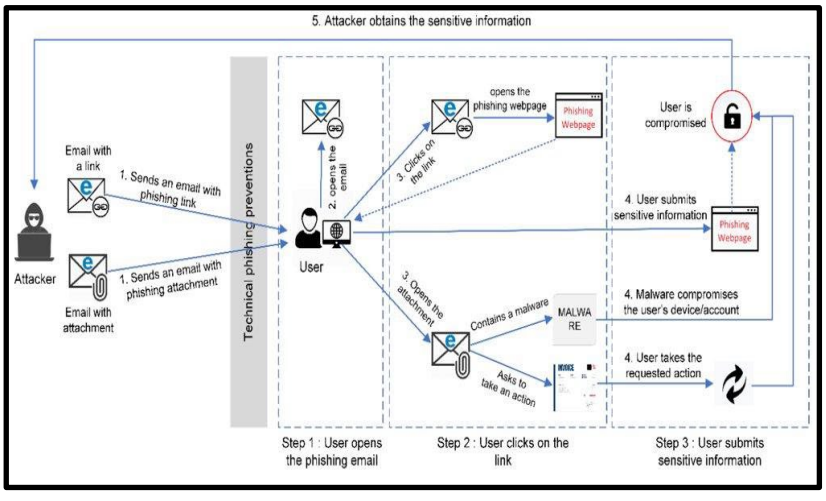


Figure 1: Phishing Example

5.1 Educating Users

Many users are not well-versed in basic browser and network security practices. This gap in knowledge can lead to risky behaviours, such as visiting unsecured websites or downloading unsafe files, which can expose users to various cybersecurity threats beyond phishing (Zieni, 2023). The primary goal of Cybersafe is to educate users on cybersecurity, equipping them with the knowledge and tools to better protect themselves online. By presenting information in a user-friendly format and providing personalized feedback based on each user's behaviour, Cybersafe aims to bridge the knowledge gap and help users develop safer online habits.

One of the key areas of focus is enabling users to recognize the signs of phishing emails. Cybersafe offers a detailed overview of common phishing tactics and provides real-time alerts when potential phishing attempts are detected. Furthermore, Cybersafe seeks to instil basic browser security methods as habits in users. Through interactive questionnaires and recommendations, users learn to identify unsecured websites, understand the importance of regularly updating browser software, and adopting secure practices like using strong, unique passwords. By addressing these critical areas of cybersecurity, Cybersafe is striving to empower users with the knowledge and habits they need to navigate the digital world safely and confidently.

5.2 The Gamified Approach

Cybersafe is a mobile application designed to help users become more aware of phishing attempts. It provides easy-to-understand and actionable information to guide users in recognizing and avoiding various types of cyber threats.

One of the unique aspects of Cybersafe is its use of gamification to teach users about various attacks. By turning learning into a game, Cybersafe makes the process of cybersecurity education more engaging and effective. Users can test their knowledge, track their progress, and even compete with friends to see who can achieve the highest Phishing Score.

The gamification component for Cybersafe is inspired from Yu-kai Chou’s Octalysis Framework (Chou, 2019). The Core Drives from the Framework that are primarily relevant to Cybersafe are Accomplishment, Empowerment, Avoidance and Unpredictability. The satisfaction and feeling of achievement by correctly identifying phishing markers and potential web vulnerabilities enable the user to feel empowered. The Black Hat core drives of Avoidance and Unpredictability also serve as strong drivers for users to engage and be motivated about solving potential phishing traps and enable them to feel a sense of control over their online safety requirements.

Cybersafe features a user-friendly interface that simplifies complex information about phishing. It breaks down technical jargon into understandable terms, making cybersecurity education accessible to users of all backgrounds and skill levels.

To ensure that users receive the most relevant and accurate information, Cybersafe uses comprehensive and up-to-date databases of known phishing websites and email addresses. These databases are continuously updated, providing users with real-time protection against emerging threats.

In addition, Cybersafe enables push notifications that alert users of potential phishing threats instantly. These real-time alerts give users the ability to act quickly, minimizing the chance of falling victim to phishing attacks.

Overall, Cybersafe provides a comprehensive and user-friendly solution to help individuals better understand and navigate the ever-evolving landscape of cyber threats.

6. Cybersafe

6.1 Operational Technology

Android Studio forms the primary development environment for the Cybersafe app. It's an industry-standard tool, offering a comprehensive suite of features to design and build applications for the Android operating system. Its range of developer tools and efficient code editor streamline the app-building process, making it easier to integrate with Android's wide array of features.

Apache Tomcat is an open-source implementation of various Java technologies that serves as the back-end server for Cybersafe. Tomcat's reliability, configurability, and strong community support make it a popular choice for building and deploying enterprise-grade applications. It handles the logic and server-side processes of the application, ensuring the smooth operation of Cybersafe's features.

Java 8 SDK is used for back-end programming in Cybersafe. Known for its versatility, robustness, and powerful feature set, Java 8 allows the application to effectively handle complex tasks. These tasks include analysing incoming data, detecting potential threats, and generating insightful educational content for the users.

Data storage in Cybersafe is managed by MySQL Database, a widely used and reliable relational database management system. MySQL facilitates efficient data storage, retrieval, and management. Its scalability and performance are critical in enabling the application to access and analyse vast amounts of data rapidly, an essential factor in detecting social engineering threats.

Servlets are employed in Cybersafe for handling requests and responses within the web environment. These server-side Java programs allow the application to extend its functionality, dynamically handling interactions with the user and ensuring efficient communication between the client and server.

Finally, Cybersafe utilizes RESTful APIs, with data structured in JSON for client-server communication. This technology promotes lightweight, stateless, and cacheable communication, boosting the overall performance and scalability of the application. Furthermore, RESTful APIs simplify the integration process with other systems and services, augmenting Cybersafe's versatility in combating social engineering attacks. We now discuss the specific operations provided by our Cybersafe application.

Cybersafe provides an array of functions designed to educate, warn, and protect users from various cybersecurity threats. The app uses a questionnaire format to assess user awareness and vulnerability to phishing, browser security, SMS phishing, and network security. These questionnaires provide valuable insights into user behaviours and facilitate personalized recommendations. The individual app functionalities are outlined below:

6.1.1 Phishing email detection

- *Analyse Links:* Evaluates the links within emails to determine if they lead to known phishing sites.
- *Sender Requests PII:* Identifies requests for Personally Identifiable Information (PII) from unfamiliar senders.
- *Expectation:* Checks if the email aligns with the user's expectations or if it's unexpected and possibly suspicious.
- *Malicious Attachments:* Scans email attachments for known malicious code or signatures.

6.1.2 Browser security

- *Public WIFI Network:* Alerts users when connected to public or unsecured WIFI networks.
- *Secure Site:* Checks for HTTPS in URLs to ensure the site encrypts user data.
- *Typos in URL or Page:* Identifies inconsistencies in URLs that could indicate a fake site.
- *Site Structure:* Analyses the overall structure of a site for signs of phishing.
- *Ads:* Monitors excessive or intrusive ads that might indicate a compromised or malicious site.
- *Browser Permissions:* Manages and monitors permission requests from websites to prevent unauthorized access to personal data.

6.1.3 SMS phishing

- *Known Number:* Compares incoming SMS from numbers with known contact information.
- *Requests PII:* Detects requests for PII within SMS messages.
- *Secure Site:* Verifies that links within SMS lead to legitimate and secure sites.
- *Legitimate Site:* Checks for signs of legitimacy, such as contact info and privacy policies.

6.1.4 WIFI security

- *Password Protected:* Ensures that the WIFI network is password protected.
- *WPA2/WPA3 Encryption:* Verifies the use of secure encryption protocols.
- *Known Users:* Monitors for unknown devices connected to the network.
- *VPN Use:* Encourages the use of Virtual Private Networks (VPNs) when appropriate.
- *Sensitive Info/Accounts:* Offers guidance on securing sensitive information and accounts when using public or shared networks.

Cybersafe's functions reflect a comprehensive approach to cybersecurity, offering robust protection across multiple domains. Its interactive and user-friendly design enables users to learn and apply cybersecurity best practices in their daily online activities, significantly reducing their risk of falling victim to cyber threats.

6.2 Operational Process for Cybersafe

Cybersafe's methodology starts with broad, general questions for each of its four categories: Email Phishing, Browsing Security, SMS Phishing, and Network Security. These questions establish a baseline understanding of the user's behaviours and potential vulnerabilities in each category. Each general question is then followed by a series of more targeted questions, which delve into specific behaviours and scenarios to gain a deeper understanding of the user's online habits.

The responses to both general and targeted questions are scored and combined to detect the likelihood of the user falling victim to phishing, spam, insecure browsing, or insecure network practices. This Phishing Score offers a quantitative measure of the user's risk level, providing an immediate insight into their current cybersecurity posture.

Based on the user's responses to both general and targeted questions, Cybersafe provides personalized recommendations. These recommendations address the specific behaviours and habits identified during the questionnaire process and provide the user with actionable advice on how to improve their online security. This approach ensures that the advice provided is relevant and helpful to each individual user, promoting a more effective learning experience and facilitating the adoption of safer online habits.

Through this combination of general and targeted questioning, scoring, and personalized recommendations, Cybersafe offers a comprehensive and personalized approach to cybersecurity education. The final operational database schema for Cybersafe is given in the appendix (Figure 2).

6.3 App Output

Cybersafe is designed to be a practical and dynamic tool that users can incorporate into their daily digital routines. When you receive an email, especially from an unfamiliar sender, use Cybersafe's phishing email detection feature. This tool analyses email for signs of phishing, such as unexpected requests for personal information, suspicious attachments, and unusual links. Like its email function, Cybersafe can also be used when receiving an SMS text message. It can detect messages from unknown senders, requests for personal information, and links to potentially dangerous sites.

Before browsing a new site, use Cybersafe's browser security function. This feature checks the site for secure connections (HTTPS), potential typos in the URL that could indicate a phishing attempt, and other signs of a secure and legitimate site.

When connecting to a new Wi-Fi network, Cybersafe's Wi-Fi security function can verify whether the network is password protected and using secure encryption protocols. It also encourages the use of a VPN when appropriate and offers guidance on securing sensitive information and accounts.

Cybersafe utilizes a unique methodology in the form of questionnaires to assess a user's vulnerability to different types of phishing and security threats. These questionnaires focus on four key areas: Email Phishing, Browsing Security, SMS Phishing, and Network Security. Each questionnaire is a mix of broad and narrow questions designed to gauge a comprehensive understanding of the user's security stance.

Through answering these questionnaires, users assist Cybersafe in grasping their online behaviours, potential threat exposure, and their comprehension of secure digital practices. This information is crucial in helping Cybersafe derive a "Phishing Score" for each user, indicating their overall risk of succumbing to phishing attacks.

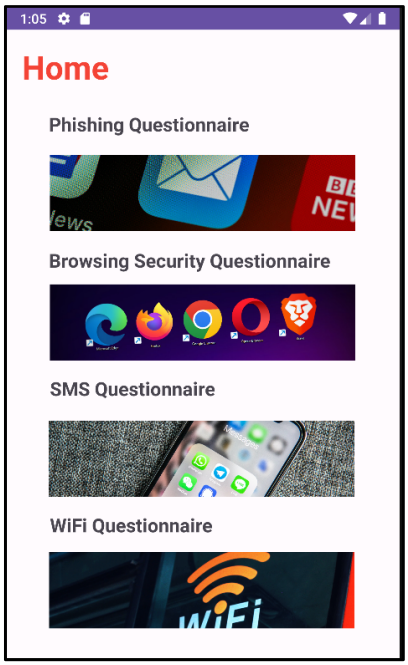


Figure 3: Main menu of mobile application, to select type of questionnaire

Each questionnaire is structured around a series of questions with three potential responses: "Yes", "No", or "Maybe". This allows users to give nuanced responses, effectively capturing their habits and potential vulnerabilities. When a user opts for "Maybe", they are encouraged to provide more context, giving further insights. Once completed, the questionnaire responses are evaluated to calculate the user's Phishing Score. A higher score represents a higher risk level.

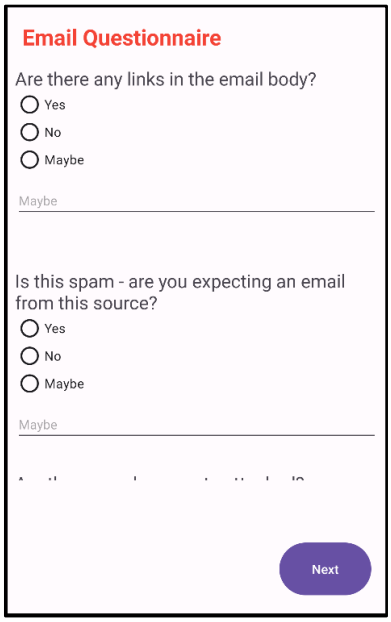


Figure 4: Example questionnaire for Browser Security

Beyond generating the Phishing Score, Cybersafe also offers personalized recommendations based on the user's questionnaire responses. These recommendations are specifically tailored around questions answered with "Yes", indicating potential security weak spots. The recommendations aim to educate users about these vulnerabilities and provide actionable steps to mitigate them.

This methodology of intertwining interactive questionnaires with personalized feedback allows Cybersafe to stand as an effective tool for increasing user awareness about social engineering threats and promoting safer online behaviours.

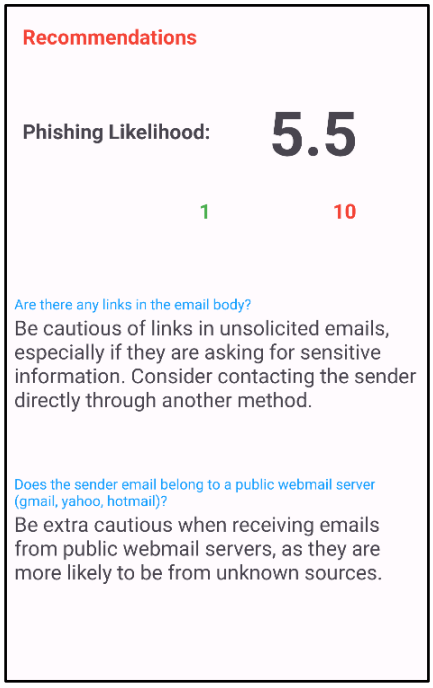


Figure 5: Example recommendations for Browser Security questions

With regular use of Cybersafe, users will enhance their ability to identify potential phishing attempts and other cybersecurity threats on their own. By integrating Cybersafe into daily online activities, users can gain practical experience in spotting and avoiding cyber threats, ultimately improving their cybersecurity skills and confidence.

7. Conclusion and Future Work

The lack of technological literacy among the public is a significant contributor to the growing prevalence and success of cyber-attacks. This gap in knowledge often leads to significant losses, both financial and personal, due to phishing and other forms of online fraud.

Cybersafe offers a solution to this issue by educating users about best practices in a personalized, engaging manner. By asking users about their online habits and providing specific questions and explanations in understandable vocabulary, Cybersafe makes cybersecurity education accessible to a wide audience.

As mobile devices continue to grow in popularity and become an increasingly common target for phishing and other cyber-attacks, it's crucial that phishing awareness and protection dynamics adapt to this change. By providing a mobile solution that users can easily incorporate into their daily routines, Cybersafe is helping to drive this adaptation, offering a relevant, user-friendly tool to enhance cybersecurity awareness and protection.

Cybersafe's goal is to empower individuals to navigate the digital world safely and confidently, thereby reducing the overall impact of cyber threats on our society. By making cybersecurity education a part of our daily online activities, we can turn the tide against cybercriminals and create a safer digital space for everyone.

As part of our future work, we plan to keep updating Cybersafe to include the latest phishing and web vulnerabilities as recorded by the cybersecurity community. This will ensure users obtain in-depth and relevant information about the current human-centric cyber threats.

Further, we plan to evaluate the usability, accuracy, and training effectiveness elements with user studies with community user participants. A pre and post evaluation measurement will be conducted to validate the usefulness of Cybersafe as an effective cybersecurity training tool and the extent of learning for the participants.

References

Abroshan, Hossein, et al. "Phishing happens beyond technology: The effects of human behaviours and demographics on each step of a phishing process." *IEEE Access* 9 (2021): 44928-44949.

Aleroud, Ahmed, and Lina Zhou. "Phishing environments, techniques, and countermeasures: A survey." Computers & Security 68 (2017): 160-196.

Bakarich, K. M., & Baranek, D. (2020). Something phish-y is going on here: A teaching case on business email compromise. *Current Issues in Auditing*, *14*(1), A1-A9.

Canova, Gamze, et al. "NoPhish: an anti-phishing education app." Security and Trust Management: 10th International Workshop, STM 2014, Wroclaw, Poland, September 10-11, 2014. Proceedings 10. Springer International Publishing, 2014.

Chou, Y. K. (2019). *Actionable gamification: Beyond points, badges, and leaderboards*. Packt Publishing Ltd.

Chaudhary, Sunil. "The use of usable security and security education to fight phishing attacks." (2016).

Chen, Tianjun, Dexin Kong, and Yuxin Hong. "Development and implementation of anti phishing wi-fi and information security protection APP based on android." Journal of Physics: Conference Series. Vol. 1802. No. 3. IOP Publishing, 2021.

Chorghe, Sharvari Prakash, and Narendra Shekokar. "A survey on anti-phishing techniques in mobile phones." 2016 International Conference on Inventive Computation Technologies (ICICT). Vol. 2. IEEE, 2016.

Dixon, Matt, Nalin Asanka Gamagedara Arachchilage, and James Nicholson. "Engaging users with educational games: The case of phishing." Extended abstracts of the 2019 CHI conference on human factors in computing systems. 2019.

Helmi, Rabab Alayham Abbas, et al. "Email Anti-Phishing Detection Application." 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET). IEEE, 2019.

Lungu, Ion, and Alexandru Tăbușcă. "Optimizing Anti-Phishing Solutions Based on User Awareness, Education and the Use of the Latest Web Security Solutions." Informatica Economica14.2 (2010).

Liu, Zhe, et al. "Nighthawk: fully automated localizing UI display issues via visual understanding." *IEEE Transactions on Software Engineering* 49.1 (2022): 403-418.

Luse, Andy, and Jim Burkman. "Gophish: Implementing a Real-World Phishing Exercise to Teach Social Engineering." *Journal of Cybersecurity Education, Research and Practice*2020.2 (2021): 5.

Wu, Longfei, Xiaojiang Du, and Jie Wu. "MobiFish: A lightweight anti-phishing scheme for mobile phones." 2014 23rd international conference on computer communication and networks (icccn). IEEE, 2014.

Zieni, Rasha, Luisa Massari, and Maria Carla Calzarossa. "Phishing or not phishing? A survey on the detection of phishing websites." *IEEE Access* 11 (2023): 18499-18519.

Appendix

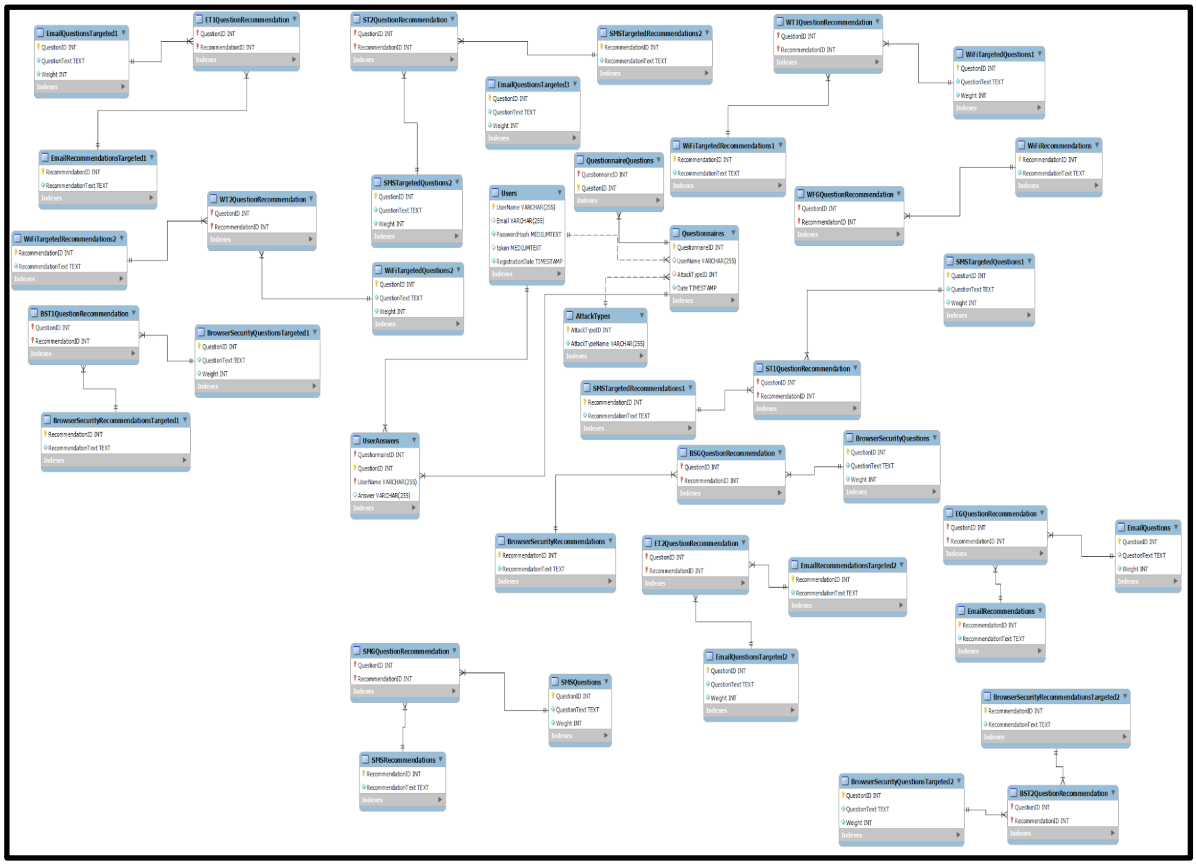


Figure 2: Final Database Schema Diagram

Analysis of Media Influence on Military Decision-Making

Bonnie Rushing
US Air Force Academy, Colorado, USA

bonnie.rushing@us.af.mil

Abstract: Information warfare challenges, including the "CNN Effect," are increasing in quantity and complexity as the internet saturates modern life. Media communications, reliable or not, are constantly injected into American life, affecting opinions, decision-making, and actions—which may result in terrifying and permanent consequences when the media influences military members and government leaders. This research focuses on the immersive learning environment—an engaging, simulation-based educational experience with realistic conflict scenarios where students apply lesson objectives through hands-on activities. Instructors conduct immersive learning in the US Air Force Academy's (USAFA) Multi-Domain Laboratory. Traditionally, students were taught military strategy in a classroom where they engaged in wargaming by rolling dice using a board game. Now, the Military & Strategic Studies (MSS) Department uses hands-on, real-time wargaming with unique hardware and software to teach future leaders how to apply classroom concepts using experiential learning. Wargames are critical for meeting the priorities of the Joint Chiefs of Staff—specifically, preparing leaders to conduct joint operations in all domains by integrating experiential learning. The guidance explains that "curricula should leverage live, virtual, constructive, and gaming methodologies with wargames… to develop deeper insight and ingenuity." The wargame with media provides external influences like news and information from various sources, which may influence students' decision-making process, resulting in "The CNN Effect." "The CNN Effect" is a short-hand way to explain how news reports drive government leaders' responses, including military decision-making. This project builds upon existing research uniquely because this study examines decision-making from multiple leaders based on the same simulated media in USAFA's immersive learning environment. This research uses mixed methods analysis to explore how students act based on the wargame media inputs and compare course outcome differences after receiving enhanced Information Operations education or when exposed to different quantities of broadcasts. The Information Warfare (news literacy) module enhances education on the media's influence on decision-making and trust for students. Students learned to identify and research information sources, analyze biases, opinion versus fact-based reporting, and compare news reports from various outlets.

Keywords: Media, Disinformation, Information Warfare, News iteracy, Propaganda, Wargaming

1. Introduction

This paper covers information warfare and media influence research conducted with cadets and military members in a US Air Force Academy (USAFA) wargame. The project meets two primary goals relating to wargame media broadcast influences on decision-making. These goals are to analyze the effects of:

- Different quantities of simulated media broadcasts.
- Additional information warfare/news literacy lesson materials.

1.1 Goal 1: Investigate Effects of Varied Quantities of Fictional Media Broadcasts

This project investigates the effect of fictional media in the Military and Strategic Studies (MSS) immersive learning environment. There are realistic media injects in the immersive learning environment of USAFA's core course, "Airpower & Joint Operations Strategy," MSS 251. The MSS immersive learning environment is an engaging, simulation-based educational experience with a realistic conflict scenario where students apply lesson objectives through hands-on activities, including encounters with fictional media broadcasts related to the wargame. The wargame occurs in USAFA's Multi-Domain Laboratory (MDL). This immersive learning environment is critical for meeting the priorities of the Joint Chiefs of Staff, who stressed the importance of developing practical warfighting skills for future wars, stating that leaders must be cognitively prepared to conduct joint operations in all domains, integrating experiential learning. Specifically, the guidance explained that "curricula should leverage live, virtual, constructive, and gaming methodologies with wargames and exercises… to develop deeper insight and ingenuity" (Milley, 2020). The desired deeper insight and ingenuity will develop by adding more realistic elements to the wargame. For one variable of this study, additional media broadcasts were added for a test group, revealing different wargame behaviors.

The wargame included four more news broadcasts for one section of students in the fall semester of 2022. The wargame typically contains four newsreels in standard MSS 251 sections, including three social media posts and two recruiting videos. These media injects create a backdrop for the wargame scenario and play on large video screens inside the MDL.

Existing research indicates that the media may compel policymakers to decide to take (or not take) military action during widely publicized humanitarian crises, violent conflicts, and other warfare events. (Doucet, 2018.)

In the unique environment of USAFA's wargame, participants make decisions similar to strategic actions from leaders as they choose to move military units, engage in kinetic or non-kinetic attacks, play defensively, work with state and non-state actors, and make other joint military decisions. This environment provides a unique opportunity to observe participants acting in a controlled setting, examining decision-making from multiple leaders (in this case, future military officers) with realistic, simulated media.

1.2 Goal 2: Investigate Impacts of Added Information Warfare/News Literacy Education

The second aspect of this project includes investigating the impact of a robust Information Operations module, enhancing education on information warfare's influence on decision-making and trust for students at USAFA in the spring semester of 2023. The lesson plan included instruction on media literacy and trust, current news and social media trends, and additional information warfare concepts. Media literacy helps people understand and judge information sources' credibility (GCF, 2021). Cadets learned to identify and research information sources, analyze biases, distinguish opinion versus fact-based reporting, and compare news reports from various outlets. Cadets also learned about echo chambers and polarization in the added module.

An echo chamber is an environment where people only encounter information or opinions that reflect and reinforce their own. This can distort a person's perspective; they have difficulty considering opposing viewpoints and discussing complicated topics. Confirmation biases partly fuel these perspectives, which is the tendency to favor information that reinforces existing beliefs (GCF, 2021). The USAFA environment may create a small-scale echo chamber and polarize student beliefs (and subsequent decision-making). Partisan echo chambers or "homogenous clusters" may exist in USAFA because students are a unique population—cadets of similar ages living in close quarters for years. Cadets may cling to their beliefs and confirmation biases, polarizing their trust (more extreme trust or distrust) of sources with additional media exposure to media from the wargame adversaries. (Guess, 2020). The cadets view simulated media that may be propaganda or information from suspicious sources.

Additionally, cadets learned in the information operations module that some influence operations deliberately mislead adversaries and cause them to make decisions or take actions based on falsehoods. The definition of "influence" includes the power to change or affect someone or something, and the data are considered by the influenced when making decisions.

Furthermore, they learned that media is created with different production values; production value includes the technical elements of a production, such as lighting, décor, speaker skills, and audio/visual qualities that increase audience appeal. These topics fall under the umbrella of information operations.

2. Literature Review

Wargaming is defined as "Wargames are representations of conflict or competition in a synthetic environment, in which people make decisions and respond to the consequences of those decisions. The Course of Action wargaming is a conscious attempt to visualize the flow of the operation, given joint force strengths and dispositions, adversary capabilities and possible courses of action, the operations area, and other aspects..." (Munsch, 2020). Wargaming stimulates students' thoughts about military operations, ideas, insights, and more.

The Joint Chiefs of Staff Vision and Guidance for Professional Military Education and Talent Management stresses the importance of developing practical warfighting skills for future wars. Warfighter skills include cognitive preparation to conduct joint operations in all domains, incorporating experiential learning. Specifically, the guidance explains that "curricula should leverage live, virtual, constructive, and gaming methodologies with wargames and exercises… to develop deeper insight and ingenuity" (Milley, 2020).

Wargames are essential for the education and development of leaders and military personnel. Learning objectives can be accomplished in a hands-on, experiential learning environment where students apply classroom concepts in an engaging, real-time environment for deeper learning. Wargames are employed in professional military education and for higher learning in numerous subjects, particularly social sciences.

Leaders must transform their learning and teaching to meet future national security challenges in the strategic competition environment. Though challenges exist for assessment strategies for educational wargames, growing research explores design, growth, testing, and more (Kuehn, 2021). Department of Defense leadership prioritizes the service members' education and skill development. They stress that experiential learning, leveraging live, virtual wargaming, and exercises develop the insight and ingenuity needed for joint operations in multi-domain warfare (Milley, 2020).

Modern challenges include information warfare, where friends and foes consistently send messages, push their narratives, have biases, and appear in social media, television, newspapers, radio, and more. Sometimes, information operations are used as open-source intelligence and may affect leaders' decision-making, leading to different actions and effects. The internet has amplified the use of misinformation, disinformation, and propaganda. People base many decisions on the internet instead of professionals and experts, allowing open-source media to alter their opinions and influence choices by impacting decisions (Power, 2011).

Awareness of propaganda and trust levels of media sources are essential for students to understand. JP 3-61, Public Affairs (PA) stresses that public affairs reporters tell the truth. "PA personnel will release only accurate, fact-based information. The long-term success of PA activities depends on the integrity and credibility of officially released information. Deceiving the public undermines the perception of legitimacy and trust in the Armed Forces. Accurate, credible presentation of information leads to confidence..." (JP 3-61, p. I-7.) Media trust is based on legitimate/authentic, *reliable*, credible, fact-based, and accurate news. Media perceptions become a person's reality and may lead to decisions and actions. Military deception information operations can deliberately mislead adversaries and cause them to make decisions or take actions based on falsehoods.

Some leaders respond to media pressures, while others resist external pressures. Specific agendas may be pushed through open-source media outlets to sway decision-makers on public policies and politics. (Koch-Baumgarten, 2010). "Fake news" and other propaganda found in social media posts and other media may affect decision-making regarding political elections. (Guess, 2020). Digital media literacy can help people understand the source of media and biases and judge the credibility of sources. (GCF, 2021).

The variable factor tested in the course is the "CNN Effect," described by Lyse Doucet's article. Lyse Doucet's article explains the CNN effect and its application in recent international conflicts. It also explains how the extent to which people are affected by the media varies based on environmental factors, individual decision-makers, and more (Doucet, 2018). Robinson's book, "The CNN Effect: The Myth of News, Foreign Policy, and Intervention," explores the history of the CNN Effect and the relationships between media and decision-making related to armed military intervention. The book describes multiple conflicts in recent history and their media coverage, exploring how the CNN Effect may have influenced government leaders to act. Robinson's study also includes the 'policy-media interaction model' that contributes to understanding media-state relations and influence (Robinson, 2006).

3. Research Objectives and Hypotheses

3.1 Research Questions (RQ)

To what extent do fictional media injects influence student decision-making (The CNN Effect) in the immersive learning environment?

To what extent do additional wargame news transmissions affect cadet trust in fictional media sources?

How does the added Information Operations module affect student decision-making?

3.2 Hypotheses

There will be evidence that the CNN Effect exists in the MSS 251 immersive learning environment and affects students' wargame decision-making and actions. (RQ 1)

There will be evidence that additional fictional news media transmissions (included only in the fall 2022 semester) will polarize students' trust (more extreme trust or distrust) of wargame information sources. (RQ 2)

There will be evidence that the robust Information Operations module will improve students' self-reported understanding of information warfare compared to standard student groups who do not experience the module, leading to more informed trust or distrust of media communications. (RQ 3)

4. Materials and Procedures

All questionnaires and other data collected were anonymous, voluntary, and provided during standard MSS 251 class meeting times at the end of academic semesters. Questionnaire feedback typically took less than five minutes to complete. *Instructors sent links for the Google Form questionnaire out over email, Microsoft Teams messaging, or provided during class time. Bonnie Rushing (PI (Principal Investigator)) was the only person accessing the questionnaire data, ensuring submissions contained no personally identifiable information and*

were saved securely. This study permits only one questionnaire submission per student. The solicitation text for the questionnaire stated, "This is an anonymous and voluntary questionnaire based on your experiences during MSS 251. Your personal data will not be collected. The questionnaire responses are for academic research purposes only. This questionnaire takes less than five minutes to complete. Thank you." Two or more class sections completed all questionnaires to increase validity. All data collected from all sources were used (full sample size).

4.1 Spring 2023 Participants

In the spring 2023 semester, cadets responded to a questionnaire indicating how fictional media influenced wargame decisions. The PI compared feedback from two groups: a *Control Group* and an *Information Module Group*.

Control Group (n = 40): Standard MSS 251 course content. No additional Information Operations module was provided. *Information Module Group* (n = 35): These students received the additional Information Operations module and the original MSS 251 content.

The Information Operations module provided greater detail on media influence for the test group. This module was added to MSS 251's standard information operations lesson. The module was a 20–25-minute addition within already approved contact hours.

Module content included news literacy, current news sources, narrative comparison, details, and history of "The CNN Effect," confirmation bias, open-source intelligence in contemporary government operations, Public Affairs guide to media, fact-based, trustworthy communications, and examination of some wargame media materials.

The PI collected and assessed control and Information Operations Module group participant reflections. This questionnaire solicits student feedback regarding wargame decision-making, information warfare influences, and trust. It also asks students about what makes media sources trustworthy or not trustworthy.

Students received the questionnaire near the end of the 2023 spring semester. The questionnaire data collected was saved in different link versions for each group. The rest of the course and wargame remained the same.

4.2 Fall 2022 Teacher Archival Course Feedback

Archival course data from the 2022 fall semester contains teacher perceptions on wargame media injects. Teachers also described the impacts of open-source intelligence on decision-making for military leaders.

Data was collected at the end of the fall semester of 2022 from four teachers in MSS 251. In this archival questionnaire, teachers provided examples of both “good” and “poor” wargame decisions and to what extent they perceived media influence on students' decision-making during the wargame.

4.3 Spring and Fall 2022 Student Archival Data

Archival course data contains student feedback after adding more wargame media for one group. Test group (n = 22), control group (n = 64).

Intervention: Three additional media injects were added to the wargame for one section. In standard MSS 251 sections, fictional media exists with fewer transmissions. The transmissions included fictitious social media posts, military and insurgency recruiting videos, and newsreels. The added media corroborated the original materials.

Procedure: Both groups completed MSS 251 with the same course content in 2022, but only the test group received additional media transmissions. Cadets responded with perceptions on wargame media injects, reflecting on the extent to which they trusted different media sources' content and how media affected their decision-making.

5. Results

5.1 RQ 1. To What Extent do Fictional Media Inject Influence Student Decision-Making (The CNN Effect) in the Immersive Learning Environment?

Wargame media influenced student decision-making, according to reflections from both teachers and students. As trends developed for this *grounded theory* research, data revealed the extent to which fictional media influenced participant decision-making.

Quantitatively, ~87% of students (of 86 cadet responses) reported that media injects affected their decision-making during the wargame.

Instructor participants explained that students should consider all information available (including media) and identify disinformation when making decisions. Data indicated that all instructors (4 of 4) perceived media influence on cadet wargame decision-making. One teacher indicated that media completely affected the students' decision-making during the wargame.

Cadets believed foreign media revealed other actors' capabilities, motivations, and objectives. 35% of cadet responses indicated that foreign media broadcasts could aid decision-making by analyzing foreign government communications as a data point and open-source intelligence. For example, one participant stated that foreign media provided "…an opportunity to gain perspectives of the adversary and their motivation for fighting. It helped with understanding how the adversary would fight and gave an insight into their capabilities."

Cadets indicated they used American and allied media as intelligence sources 30% more often than foreign media. They stated that they relied on American and allied media as open-source intelligence or data points for their decision-making in the wargame, that they trusted the American news broadcasts, and that these media broadened their situational awareness and used its information to prioritize military operations. Cadets believed that the US and allied media contained biases, but gave participants insight into the local conditions, created concern for the Japanese people (simulated humanitarian crisis), motivated them to work hard to help their partners, and caused confirmation bias by positively reinforcing their thoughts and decision-making. A participant stated, "The American and allied media affected my decision-making, especially HNN, because I made decisions that responded to the events unfolding in the news channel and believed most of them to be true."

Students questioned the trustworthiness of foreign communications and narratives. 36% of student responses highlight that foreign media broadcasts in the wargame were false, propaganda, and disinformation. An example remark stated, "Most of the foreign media was taken as false and used to discern motive reasoning for propaganda."

Cadets noted that American and allied media can also be used as propaganda to steer the narrative in favor of democratic values. 5% of students believed that media is essential for the promotion of American values, support for military operations, building alliances, and combating adversarial information operations. Only 19% of participants commented that foreign media had no influence, and 16% stated that US/allied media "did not influence" them in any way.

This data also revealed that the foreign media influenced 6% of cadets to operate more aggressively in the wargame. These cadets indicated that the adversarial media caused them to make bolder decisions, trigger an emotional response to retaliate, and employ military offensives against the opponent. "…The foreign media influenced me to want to fight back more and make stronger/bolder decisions." The adversary's media broadcasts "strengthened my resolve against the [adversary] forces and liberating [allies] from their propaganda [expletive.]" American and allied media did not have this effect.

5.2 RQ 2. To What Extent do Additional Wargame News Transmissions Affect Cadet Trust in Fictional Media Sources?

Grounded theory explored emerging themes from students' comments regarding media influence, with and without additional fictional media transmissions. (Table 1).

Test Group: Students who received extra corroborating media broadcasts in addition to standard content (n = 22). *Control Group*: Students who received only standard MSS 251 wargame content; no additional media (n =

Table 1: Statistical analysis of cadet responses based on quantity of media transmissions.

	Control Group (standard content)	Test Group (additional injects)
Injects did not affect their decision-making	10.9%	18.2%
Trusted American/allied social media	75%	77.2%
Trusted American/allied social media	59.4%	63.6%
Distrust in foreign newsreels	65.6%	86.3%
Distrust in foreign social media	75%	81.8%

5.3 RQ 3. How does the Added Information Operations Module Affect Student Decision-Making?

This research compared data from the two groups of students, detailing the extent to which trust (JP 3-61, p. I-7) and decision-making are influenced by media—with or without exposure to enhanced information operations education.

Compared to the control group (n = 35) of students who completed the standard MSS 251 course, the *Information Operations module (test) group, n = 25* reported that they *"thought about media injects" 32.6% more often, "discussed the media with classmates" 44% more often, considered "the trustworthiness of each media source" 34.9% more often*.

The additional Information Operations education impacted the following student responses, specifically to the wargame. The *control group "did nothing" related to the media 17.7% more often*. The *Information Operations module group "conducted simulated humanitarian assistance/disaster recovery" 30.3% more often* and *"decided to work with the fictional rebel group" 37.7% more often.*

Humanitarian assistance, disaster recovery, and security force assistance operations are taught as "military operations other than war" in the MSS 251 course. Students were more likely to simulate these operations >30% more often after receiving a robust information operations module than their control group counterparts.

Tables 1 and 2 represent the discovered trends in support of qualitative grounded theory from the test group (extra Information Operations module) and the control group students.

The first qualitative prompt asked cadets, "What made media seem trustworthy?" 40% of both groups of students, regardless of added Information Operations education (test group), reported that the reputation of media sources added to the trustworthiness of its communications. Examples of these comments included "lead with reputable source name" and "official, major news network."

Table 2: Statistical Analysis of hand-coded cadet trend comments on media trustworthiness

What made the media seem trustworthy?		
	Control	Test - Extra IO Module
Source reputation	40%	40%
Production quality	11%	12%
Data-based/unbiased	20%	32%
What made the media seem **NOT** trustworthy?		
	Control	Test - Extra IO Module
Source reputation	17%	40%
Production quality	26%	24%
Not data-based/biased	17%	36%

The test group's remarks revealed that they believe *media trustworthiness relies on unbiased, data-driven reporting 12% more often than the control group.* Some examples of these responses included "factual information rather than political rhetoric" and "research and statistics that reduce bias."

The second qualitative prompt asked, "What made media seem NOT trustworthy?" *Both groups were similar (+/-2%) in their reporting of production quality* for this question. Some responses coded this way included "disconnected transmission, sounded spotty," and "anything spoken or written that sounded like broken English." The control group responses (26%) and the test group (24%) almost doubled for this prompt. This result reveals that *cadets felt more strongly about poor production quality as an indicator of suspicious media.*

The *test group reported the source reputation 23% more often as the reason to NOT trust media*. Examples include "if the platform was uncertain (Twitter, social media) or the source was Chinese," "communications from the adversary – propaganda," and "when it came from just individuals or those who held no reputation." The

control group's responses only revealed this data in *17% of the responses, lower than their data for production quality (26%).* This result shows that *without added Information Operations education, students were more concerned about poor production quality rather than considering the source of suspicious media communications*. There was a 9% statistical variance between these trends.

In contrast, the test group considered the source reputation at the same frequency (40%) in both trustworthy and untrustworthy media; *the test group considered the source reputation more so than any other trends for both prompts.*

The test group's remarks revealed that they believe *media seems untrustworthy if biased and not data-driven, reporting this trend 9% more often than the control group*. Some examples of these responses included "lack of statistics, or vague, unsupported studies," and "if they were clearly biased one way or another instead of taking a neutral stance."

6. Conclusion

Information warfare challenges, including the *CNN Effect*, are growing as the internet saturates our lives. Americans are increasingly exposed to social media and other broadcasts with the expansion of smartphones and available network venues. Media communications, reliable or not, are injected into our lives and can cause psychological manipulation by swaying our opinions, decision-making, and actions. These influences could have terrifying and permanent consequences when military members are affected.

USAFA's cadets will be the nation's leaders. Analyzing their decision-making and wargame actions based on media makes it possible to reveal information warfare's influence on the next generation. In 2022, roughly 87% of cadets reported that media injects affected their decision-making during the wargaming scenario, and four out of four instructors perceived media influence on their students' actions. Social media and news broadcasts undoubtedly influence the thoughts and decision-making of military officials to varying degrees.

Participating instructors explained that media communications and open-source intelligence might reveal truths and serve as data points. However, leaders must understand and think critically about source reliability, disinformation, and possible adversarial information operations. Similarly, 35% of cadet responses indicated that foreign media broadcasts could aid decision-making by analyzing foreign government communications as a data point and open-source intelligence.

One participating instructor warned that foreign media should not be used to drive major decisions. This warning may only be apparent to some cadets. About 6% of participants reported that the media *totally* affected their decisions. If media influences 100% of a participant's decisions, there is a vulnerability to malicious influences and exploitation. Persuasive messages can originate in both foreign and domestic media.

Data also revealed that the fictional foreign media influenced 6% of cadets to operate more aggressively in the wargame. These cadets indicated that adversarial communications triggered an emotional response to retaliate, causing them to make bolder decisions and employ military offensives against the opponent.

Participants stated, *"…The foreign media influenced me to want to fight back more and make stronger/bolder decisions,"* and that the adversary's media broadcasts *"strengthened my resolved against the [adversary] forces and liberating [allies] from their propaganda [expletive]."* Therefore, social media posts, radio transmissions, and news broadcasts may play a decisive role in wartime morale, aggression, and emotional conduct on the battlefield. Examples of emotional and aggressive conduct in the wargame included careless attacks on the enemy without first gathering supporting forces or intelligence, resulting in needlessly lost lives, ships, and aircraft. This is a simulated result of adversary information operations.

Data varied when members consumed increased amounts of news and social media. Participants exposed to more media broadcasts reported *more extreme trust or distrust* in all types of fictional communications, causing a *polarizing* effect. Added iterations of foreign news and social media destroyed the trust of students who saw more of it. They reported higher trust in American and allied news (2.2%) and social media (4.2%) and were more suspicious of foreign news (20.7%) and social media (6.8%).

Higher levels of news consumption created a polarizing effect; cadets felt more extreme trust or distrust of the sources than the control group. They trusted the information that fit their beliefs (US/allied) to higher levels upon exposure while rejecting alternative views such as foreign media. Media source trust levels plunged (polarized) further into extreme directions after consuming higher quantities of news content.

Similarly, data revealed that media literacy lesson materials affected media influence on participants. When populations learn about propaganda through experiential methods, they should be better equipped against harmful external influences.

Information warfare education was tested in this project to analyze its effects and possible benefits for enhanced critical thinking, mitigating adversarial psychological operations through media literacy, and for use in the MSS 251 course. Cadets who experienced the added module (test group) revealed that they believe media trustworthiness relies on unbiased, data-driven reporting 12% more often than the control group. Additionally, they thought about media injects 32.6% more often, discussed the media with classmates 44% more often, and considered the trustworthiness of each media source 34.9% more often. They actively engaged with the presented news, discussing the information sources in class and the lab more often.

The test group also conducted simulated humanitarian assistance/disaster recovery 30.3% more often and decided to work with the non-state actors 37.7% more often. These actions in the wargame require more coordination among the team and a deeper understanding of the operational environment. The test group's unique information warfare education, media literacy, and communication engagement may correlate with these human-centered missions, simulating these operations >30% more often.

The control group “did nothing” related to the media 17.7% more often and was more often concerned over poor production quality than the media source (by 9%) when considering the trustworthiness of communications. Without media literacy education, populations may trust fake news with Hollywood-level production (smoke and mirrors) value without considering the source's reputation and legitimacy. Since production quality is a proxy for lower suspicion of media and communications, adversaries may produce high-quality content to gain influence.

This evidence shows the possible benefits of the Information Operations module. Students with the added lesson material did not take all media at face value or place a high value on the smoke and mirrors of production quality over fact-based content. For competitive advantage in the growing information warfare landscape, leaders must be critical thinkers and engage with the media appropriately. Media literacy lesson materials better prepared the test group for simulated warfare, reducing the potential influence of disinformation and other external communications.

Disclaimer: *The views expressed in this document are those of the author and do not reflect the official policy or position of the USAF, DoD, or the US Government.*

Acknowledgements

I want to thank the US Air Force Academy, the study participants, the Military and Strategic Studies Department, and the Multi-Domain Laboratory team for their support of research, scholarship, and growth. Additionally, I would love to thank USAFA’s Mastery of Teaching Certificate Community for helping shape this study: Dr. Lauren Scharff, Dr. Marc Napolitano, and all individual members. Thank you to Dr. Ethan Eagle and Dr. Paul Bezerra for impromptu help. And, of course, to my mom, Kim Natcher, who has always been my biggest fan. I love you.

References

Bartels, E. (2020) “Building Better Games for National Security Policy Analysis,” [online], *RAND Corporation*, https://www.rand.org/pubs/rgs_dissertations/RGSD437.html

Bartels, E. (2021) “Wargames as an Educational Tool,” [online], *RAND Corporation*, https://www.rand.org/blog/2021/02/wargames-as-an-educational-tool.html

Golby, J. (2020) “Want Better Strategists? Teach Social Science,” [online], *War on the Rocks*, https://warontherocks.com/2020/06/want-better-strategists-teach-social-science/

Guess, A. and Lyons, B. (2020) “Social Media and Democracy.” *Social Media and Democracy*, 24 Aug, pp. 10–33.

Koch-Baumgarten, S., and Voltmer, K., editors. (2010) *Public Policy and the Mass Media: The Interplay of Mass Communication and Political Decision Making*. Routledge, London, United Kingdom.

Kuehn, K. (2021) “Assessment Strategies for Educational Wargames.” *Journal of Advanced Military Studies*, vol. 12, no. 2, 1 Oct. 2021, pp. 139–153.

Lyse D. (2018) "Syria & the CNN Effect: What Role Does the Media Play in Policymaking?” *Daedalus, 2018*, 147 (1): 141–157.

Mayville, W. (2015) "Joint Publication 3-61," [online], US Department of Defense Public Affairs, https://www.jcs.mil/Portals/36/Documents/Doctrine/pubs/jp3_61.pdf.

Milley, M. (2020) “Developing Today's Joint Officers for Tomorrow's Ways of War,” [online], *Joint Chiefs of Staff Vision and Guidance for Professional Military Education and Talent Management*, Joint Chiefs of Staff,

https://www.jcs.mil/Portals/36/Documents/Doctrine/education/jcs_pme_tm_vision.pdf?ver=2020-05-15-102429-817

Munsch, S. (2020) "Joint Publication 5-0 Planning," [online-] *JP 5-0. Joint Planning*, Joint Chiefs of Staff, https://www.jcs.mil/Doctrine/Joint-Doctrine-Pubs/5-0-Planning-Series/

Power, D., and Phillips-Wren, G., (2011) "Impact of Social Media and Web 2.0 on Decision-Making." *Journal of Decision Systems*, vol. 20, no. 3, pp. 249–261.

Robinson, P. (2006) *The CNN Effect: The Myth of News, Foreign Policy, and Intervention*. Routledge.

The Goodwill Community Foundation, GCF. (2021) "Digital Media Literacy: The Blur between Facts and Opinions in the Media," [online], *GCFGlobal.org*, https://edu.gcfglobal.org/en/digital-media-literacy/the-blur-between-facts-and-opinions-in-the-media/1/

Building Cybersecurity Capacities in Zambia's Business Sector: Guideline for SMEs

Goni Saar[1] and Rabelani Dagada[2]
[1]The Da Vinci Institute for Technology Management, Zambia
[2]Graduate School of Business Leadership, University of South Africa, South Africa

Gonisaar2011@gmail.com
dagadr@unisa.ac.za

Abstract: This research explores cybersecurity awareness and implementation within Zambia's small and medium-sized enterprises (SMEs), a sector increasingly targeted by cyberattacks that lead to substantial financial losses. The study's primary aim was to enhance cyber awareness and develop actionable guidelines for SMEs in Zambia. Utilising an interpretivist philosophy and inductive approach, the methodology encompassed semi-structured interviews, cross-sectional analysis, and a comprehensive review of CISA, ENISA guidelines, and Zambia's Data Protection Act. Findings indicate a notable deficit in cybersecurity training and awareness among SMEs. Key concerns include inadequate data security measures, a lack of formal cybersecurity policies, and a reliance on basic tools like antivirus software. In response, the study formulated targeted guidelines, emphasising the integration of cyber awareness into SME governance and risk management. These guidelines have garnered significant interest from Zambian government entities, highlighting their potential influence on national cybersecurity policy. The study contributes theoretically by contextualising international cybersecurity standards within Zambia's unique SME landscape. Methodologically, it pioneers a Cyber Awareness Framework tailored to Zambian SMEs, underscoring the critical role of human factors in cybersecurity. Practically, the research has sparked engagement among SMEs and government bodies, demonstrating its applicability and potential for shaping policy. However, limitations include reliance on outdated demographic data and a focus on digitally enabled SMEs, potentially overlooking broader IT governance aspects and less digitized businesses. Future research should aim for comprehensive, up-to-date analysis across all SME sectors, contributing to a more inclusive and resilient cybersecurity landscape in Zambia.

Keywords: Cybersecurity awareness, SMEs, Zambia, Data protection, Cybersecurity guidelines, Financial impact of cyberattacks

1. Introduction and Background

Among the most pressing challenges of the 21st century is cybersecurity threats, which negatively affect businesses and nations worldwide. Due to the modern lifestyle that prevails, people nowadays rely on technology for their daily activities such as shopping, financial transactions, and other aspects of their daily routines (Rajasekharaiah, et al., 2020). Additionally, with the growth and popularity of social media, cybercrime has concomitantly increased.

Most government leaders, in both developing and developed nations, have embarked on providing cybersecurity because it is the key to stimulating prosperous and enhanced national security (Shafqat & Masood, 2016). However, cyber threats continue to increase despite these efforts.

In the digital age, cybersecurity has emerged as a critical issue due to the rising number of cyberattacks and data breaches. This has resulted in significant financial repercussions. The economic significance of cybersecurity is a notable factor in the digitalization of national economies. This multifaceted aspect can be examined from various angles, including the cost of cybercrime, the impact of cyberattacks on businesses, and the economic benefits associated with cybersecurity investments (Dagada, 2013).

Morgan (2022) highlights the escalating costs of cybercrime, projecting damages of 8 trillion dollars in 2023 and 10.5 trillion dollars in 2025, compared to 3 trillion dollars in 2015. Such statistics underscore the growing financial consequences of cyber threats and the urgency to address them effectively.

In the business sector, the connection between businesses and cybersecurity has become indisputable in the 21st century. Dagada (2021) emphasises the pervasive adoption of digital platforms by businesses worldwide. The fourth industrial revolution, characterized by rapid technological advancements such as artificial intelligence, the Internet of Things, and robotics, further underscores the need to address cybercrimes associated with these technological developments. With both the government and private sectors embracing digitalisation, cybersecurity is no longer an optional consideration but an imperative investment to mitigate potentially substantial losses incurred through cyberattacks (Dagada, 2014).

Cyberattacks can inflict significant damage on businesses, ranging from reputational harm and financial losses to legal liabilities. A study conducted by the Ponemon Institute found that the average cost sustained through a

breach of data for a company in 2020 was $3.86 million (Ponemon Institute, 2020). Business-related cyberattacks can have wider implications and pose a national threat to nations.

Kozak (2017) emphasises the criticality of regional economies, wherein small and medium-sized enterprises (SMEs) hold prominence in driving a country's economic development. This importance is particularly pronounced in rural areas, where SMEs serve as primary employers, fostering local economic growth (Kozak, 2017). According to the Zambia Development Agency (ZDA) (2020), SMEs account for an estimated 70% of the employed population and contribute about 20% to GDP.

Every year, Zambia is subjected to an increasing number of cyberattacks that result in losses of hundreds of millions of Kwacha (National Assembly of Zambia, 2022). Together with the rapid growth of ICT use in Zambia (ZICTA, 2021), dealing with this substantial cybersecurity issue has become a major challenge for the country's economy. The challenge of cyber security in Zambia may be attributed to businesses or companies not being sufficiently knowledgeable about cybersecurity.

Zambia has sustained significant economic impairment due to cyber threats, with the Zambia Computer Incident Response Team reporting a surge in cyberattacks, culminating in losses of over 150 million Zambian Kwacha in 2021 alone, including losses from fraudulent investment schemes (National Assembly of Zambia, 2022). Contributing to the rise of such cyber incidents is a notable deficiency in cybersecurity awareness. This is highlighted by a study on information security awareness among Zambian higher education employees, which found a lack of adequate information security training and a dearth of support from the upper echelons of management.

Zambia's legislative advancements, particularly the enactment of the Data Protection Act No. 3 of 2021, established a foundational legal framework for personal data protection through the establishment of the Office of the Data Protection Commissioner (Government of Zambia, 2021). Despite these steps, Zambia faces ongoing cybersecurity challenges, including technical skill shortages, infrastructural inadequacies, and a general lack of cyber threat awareness. A comprehensive strategy that integrates legislative efforts, capacity building, and public education, augmented by international collaboration, is essential to navigating these challenges (ZICTA, 2022).

The purpose of the study was to explore knowledge and practices of cyber security amongst individuals in the Zambian business sector, and furthermore to develop guidelines that enhance cyber awareness amongst employees.

2. Theoretical and Conceptual Framework

This research utilizes the "Cyber Security Awareness and Education" framework as formulated by Kortjan and von Solms (2014). This framework was originally designed for enhancing cyber security culture at a national level within South Africa. The framework, informed by an analysis of cyber security practices in OECD countries, proposes a five-layered structure for cyber security awareness. However, for the purpose of this research, only three layers most pertinent to the organizational context have been utilised: strategic, tactical, and monitoring.

Although this framework was intended for national use, it also offers valuable insights for the private sector. By considering a company as a "microcosm of a country", we can apply similar measures to instil cybersecurity awareness within the business sector.

The conceptual framework of this study, derived from the theoretical framework and supplemented by a concept identified through the comprehensive literature review, is built upon a selection of core concepts pertinent to the business sector. These include cybersecurity policy, responsibility, training, monitoring, and actions.

These concepts resonate with the broader definitions provided by Schatz et al. (2017) and Shaw et al. (2009), which emphasise the importance of policy, training, and individual responsibility in creating a robust cybersecurity environment.

A comprehensive review of CISA's and ENISA's guidelines for SMEs was undertaken in this study. This analysis was conducted in alignment with the previously defined concepts of cybersecurity and cyber awareness, as well as the established conceptual framework. The primary objective of the analysis is to discern which guidelines are most conducive to enhancing cyber awareness among businesses in Zambia and which are instrumental in augmenting the cybersecurity posture of Zambian enterprises. This systematic approach ensures that the

selected guidelines are not only relevant to the Zambian context but also effectively address the twin objectives of raising cyber awareness and strengthening cybersecurity in the business sector.

3. Research Methodology

The qualitative research methodology was adopted for the study. And the following data collection instruments employed: semi-structured in-depth interviews, cross-sectional tool, and document analysis. The adopted strategy was grounded in a humanistic and interpretive perspective, aiming to deeply understand human experiences. The choice of this approach was particularly effective for exploring cybersecurity awareness among professionals in Zambia's business sector, facilitated through semi-structured interviews. Additionally, document analysis provided a complementary method, helping to extract pertinent themes, best practices and legislative frameworks. This combination of techniques ensured a thorough and dependable foundation for the study's conclusions.

Population and Sampling

Employees from small and medium-sized enterprises formed the target population for the study, acknowledging the vulnerability of SMEs to cyber threats due to typically fewer resources and less stringent policies than larger firms. The research included employees across organisational levels rather than focusing on senior management alone, recognizing that cyberattacks can target any employee and that responsibility for cybersecurity spans the entire organisation.

The study employed purposive sampling, a non-probability technique ideal for qualitative research, allowing for in-depth exploration of specific traits within the participant pool. Twenty participants were chosen for their direct experience with cybersecurity in the Zambian business context, leveraging the principal researcher's industry connections to identify individuals who could provide relevant insights. This sampling method ensured that interviewees could contribute meaningfully to the research objectives based on their knowledge and experience.

Data Analysis

Thematic analysis was the primary method for analysing interview data in this study, involving a six-step process recommended by Kiger and Varpio (2020). It began with examining the interview data, coding to categorise the information, and identifying themes relevant to the research context. Coding was essential for comparing data segments and facilitating the formulation of theoretical constructs.

Interview data were analysed for semantic content, with themes cross-referenced against literature to provide comprehensive insights. Additionally, document analysis of CISA and ENISA guidelines and Zambia's Data Protection Act informed the findings, ensuring industry relevance and legal compliance. This multi-layered analysis yielded a robust and holistic understanding of cybersecurity challenges and practices in Zambia's SME sector.

4. Findings of the Study

The findings of this study are grounded in a methodical examination of three primary sources: the guidelines from CISA and ENISA specifically designed for SMEs; the nuanced thematic analysis of in-depth interviews; and a thorough review of Zambia's Data Protection Act, which is further enriched by an in-depth discussion with the Data Protection Commissioner. The structure of the research findings is sequential, with each layer contributing to a more profound understanding of cyber awareness and security in the Zambian business context. The guidelines developed from this process are presented in the appendices, tailored to enhance cyber awareness and cybersecurity measures for businesses in Zambia.

Major Theme 1: Cybersecurity Knowledge and Practices among Employees

In an increasingly digital landscape, cybersecurity is essential in seeking to safeguard assets (Gundu, 2019). Employees are a principal defence line against cyber threats (Nifakos, et al., 2021). Major Theme 1, "Cybersecurity Knowledge and Practices among Employees", explores employees' cybersecurity awareness and actions in Zambia's business sector. This theme investigates their understanding of cyber threats, engagement in preventive measures, response to incidents, and their role in organizational cybersecurity. It aligns with the study's first objective, exploring employee cybersecurity knowledge and practices in Zambia, aiming to identify gaps and training needs. Subsequent sections analyse this theme, offering a detailed view of employees' cybersecurity awareness and practices in Zambia.

Sub-Theme 1.1: Use of IT Systems and Cybersecurity

IT systems have transformed organizational operations, enhancing efficiency and innovation (Baskerville, et al., 2018), while also raising cybersecurity stakes (Slusky, 2020). The first sub-theme, "Use of IT Systems and Cybersecurity", under Major Theme 1, examines the relationship between employees' IT usage and cybersecurity. It looks at how tasks like emailing, data management, financial transactions, and business processes (Boyce, et al., 2011) intertwine with cybersecurity. This sub-theme explores employees' perceptions and integration of cybersecurity in IT usage, probing their awareness of technology and security interplay.

The focus is on whether employees see cybersecurity as inherent to IT usage and understand the risks of a lax approach. It also examines the demand for enhanced security in IT systems use. This analysis aligns with the theme's goal of dissecting Zambian employees' cybersecurity knowledge and practices. The following sections explore these aspects, revealing how employees' interaction with IT systems intersects with their cybersecurity awareness.

Integration of IT Systems in Professional Duties

IT systems are integral to professional duties, with cloud-based tools like Sage 200 enhancing financial tasks through real-time access ("I'm using an accounting package which is Sage 200 and that's connected on the cloud..."). Essential daily tools include desktop printers, computers, the internet, emails, and WhatsApp ("Desktop printers, Computers, Internet, Wifi, Emails and WhatsApp").

Dual Usage of IT Systems - Personal and Professional

IT system use often overlaps between work and personal life, with common tools like laptops, phones, and the internet serving both purposes ("...We use the laptops and the phones... Gmail. Emails we have"). Emails and the internet, while primarily for professional use, are also used for personal interactions, as indicated by participants' usage of computers for email and internet tasks ("...The computer where we receive mail..."; "Not really complex ones, but I do. Internet, emails").

Varied Approaches to IT System Usage

Individuals vary in their IT system usage, with some prioritizing security through tools like Google Drive and official emails (*"...upload PODs in the Google Drive"*), while others use a broader array of tools, including personal apps like WhatsApp for work. Examples include using internet, office tools, SAGE, and emails for various tasks (*"...financial packages, emails"*). This sub-theme also explores how Zambian employees integrate diverse IT systems into their professional tasks, focusing on the alignment with cybersecurity considerations.

Diverse Spectrum of IT Systems Usage

Professional settings utilize a broad spectrum of IT systems, including desktop printers, computers, the internet, emails, and WhatsApp, reflecting diverse technology integration (*"Desktop printers, Computers, Internet, Wifi, Emails and WhatsApp"*).

Specialized IT Tools and Applications

The use of specialized tools like NetSuite, WAN, and tracking systems illustrates the adoption of diverse technologies tailored to specific organizational needs *("...using a tracking system..."; "...our programmes, NetSuite, WAN"*). This indicates a deliberate effort to integrate technology that meets unique business requirements.

Dual Usage and Flexibility

Employees often use IT resources like laptops, emails, and payroll systems beyond office hours, merging work and personal life (*"Just the laptop, sending emails, receiving emails"*). This highlights the flexibility and accessibility of modern technology (*"We've got a payroll system, a system that captures data..."*).

Leveraging IT for Efficient Reporting

IT systems are indispensable for efficient reporting in organisations, with practices like sending reports via the internet and using software like WIN (ERP system) and NetSuite enhancing communication and decision-making (*"Sending reports, when I'm doing my reports, IT is involved. I use the internet, the computer"; "Emails"; "WIN"; "NetSuite"*).

These findings emphasise the significant role of IT systems in work efficiency and effectiveness but also the need to balance enhanced productivity with robust cybersecurity, considering the overlap of personal and professional use. The diversity and adaptability of IT system usage in the workplace, from standard to specialized applications, are highlighted, demonstrating their importance in organisational communication and data management.

Sub-Theme 1.2: Perception and Understanding of Cyber Threats

This sub-theme examines how Zambian business employees perceive and understand various cyber threats, focusing on their awareness, knowledge of different threats, and assessment of their severity and impact. It explores their awareness sources, including training, personal experience, or media, and aims to understand the mental frameworks used to interpret these threats. This understanding is crucial for developing effective cybersecurity measures and strategies to enhance awareness and prevention. It also aligns with the second research objective: identifying factors influencing cyber awareness implementation among employees. The following sections analyse how these employees view and respond to the dynamic cyber threat landscape.

Perception

Interview participants recognized phishing and malware as major cyber threats, understanding the tactics used by cybercriminals (*"Phishing malware...they extort information from you which they later use to threaten you or bribe you..."*). They demonstrated awareness of the evolving cyber threat landscape and its potential harm, emphasizing the importance of proactive security measures (*"What I know about cyber risks is that we need to protect ourselves from the cyber risks or the cybercrimes..."*).

Participants acknowledged the vulnerability of personal information in the digital sphere, highlighting the need for protective measures against data exposure (*"They mostly expose sensitive information that should not be revealed to the public"*). Concerns about hacking and the risk of unauthorized access to systems were prevalent, underscoring the necessity of strong defences (*"If you're on the internet and you see some link from a certain individual that you don't even know about, then you open that link, it might corrupt your PC or they might steal information from you..."*). Negligence was also seen as a significant threat, pointing to the need for a culture of vigilance and responsibility (*"It's negligence"*).

Furthermore, the awareness of cyber threats extended to specific organizational sectors, with particular emphasis on protecting guest information (*"So that one, I might say reviewing guest information which is confidential guest information"*) and identifying fraud as a key concern (*"Fraud, maybe"*). The vulnerability of IT infrastructure, especially a single server, was recognized as a critical issue (*"I think the biggest threat is that, we have a few, we have a single server"*). Email communication was noted as a potential risk area, particularly regarding the transmission of sensitive information (*"Information. On the information part especially because we deal with lots of emails and those emails there are some which are very sensitive"*).

Overall, participants exhibited a clear understanding of various cybersecurity threats, recognizing the need for comprehensive protective measures and strategies to mitigate these risks. This insight reflects a growing consciousness of the importance of cybersecurity in both personal and organizational domains.

Experience of and Response to Cyberattacks

Participants reported varied personal experiences with cyber incidents. Some faced cyber-attacks, while others took proactive steps like password protection to prevent threats (*"Yes, we had to put some measures, like block back pacing and put some security measures on our group, add a password to it"*). A few, particularly those with an IT background, had not experienced incidents, attributing this to their knowledge or cautious practices (*"Personally, yes, but since I have a bit of IT background, I know how to handle them"*). However, many participants did encounter cyber-attacks, highlighting their widespread and persistent nature (*"I think I've also encountered, I'll call it cyber attack..."*). These experiences emphasise the necessity of proactive and robust cybersecurity measures to effectively mitigate potential threats.

Training to Counter Cyberattacks

Participants highlighted a significant lack of cybersecurity training, with many noting the absence of formal programmes (*"No, I've never had one"; "No"; "No, we've never really had any sort of training or any sort of awareness pertaining to the risks"*). Responses varied from complete lack of training to some training that was deemed insufficient or not focused on cybersecurity risks (*"No, I would like to start from the beginning, the basics..."*).

Despite this gap, there was an eagerness among employees to learn more about cybersecurity and how to tackle cyber threats (*"Through training... You need to learn more on how to identify risks"*). The overarching theme was the absence of structured cybersecurity training in organizations (*"No, we've never really had any sort of training or any sort of awareness pertaining to the risks"; "No"; "Personally, no"*).

These findings emphasize the need for comprehensive and regular training programmes that are tailored to employees' needs and roles, highlighting the importance of enhancing knowledge and awareness to strengthen the organization's defence against cyber threats.

Response to / Countering Cyberattacks

Participants altered their behaviour after cyber incidents, becoming more cautious and less trusting online (*"Yeah, so I've learned not to trust anyone"; "If you are not sure of anything don't click on anything that we're not sure of"*). They adopted practices like changing passwords and avoiding risky online behaviour (*"Yeah like on my personal space I've changed my passwords..."; "I just don't involve myself in these malicious pages, like especially porn sites"*).

There was a strong desire for more cybersecurity knowledge, with employees emphasizing the importance of training (*"Training, training can do"; "I think it's something that we are, it's about just awareness and training, it's something that we are working on"; "I would just love to have a little bit more knowledge about cyber security"*). This underscores the need for regular training and proactive practices to foster a cybersecurity-aware culture and enhance organisational resilience against cyber threats.

Major Theme 2: Factors Influencing Cyber Awareness Implementation

This Major Theme explores the various elements necessary for the effective integration of cybersecurity awareness programmes in organisations. It covers a range of factors including leadership support, organisational culture, resource availability, and technological infrastructure, essential for fostering a cybersecurity-conscious environment. The findings in this theme highlight the key determinants for successful cyber awareness initiatives, offering insights for organizations looking to strengthen their cybersecurity position.

Cyber Risks / Threats; Awareness and Training

The sub-theme reveals that perceiving and understanding potential cyber risks and threats is fundamental to establishing a proactive cybersecurity defence (Yildirim, 2016). Integrating awareness and training is vital for enhancing organizational resilience against the evolving cyber threat landscape (Chen & He, 2013). Participants unanimously recognized the ubiquity and diversity of cyber risks, with hacking and blackmailing through hacking cited as common concerns (*"Cyber risks has advanced, people hacking emails, even your personal laptop can be hacked"; "With this advanced internet, there are so many risks. Mostly the hackers"*). The fear of cyber violence and bullying was also prevalent among respondents.

Interestingly, many participants, especially those knowledgeable in IT, had not experienced cyber incidents, attributing this to proactive security measures like strong passwords. However, some noted a lack of adequate protection for organizational systems, emphasizing the need for enhanced security measures (*"The biggest threat, if our system is not well protected..."*).

These findings underscore the importance of fostering awareness and training to empower employees with the necessary skills to effectively identify and mitigate cyber risks. Additionally, they highlight the need for improving protective measures and strengthening security infrastructure as critical steps toward a resilient cybersecurity posture.

Awareness

Awareness of cyber threats is necessary for a robust cybersecurity framework (Ifinedo, 2023). This sub-cluster explores how individuals learn about cyber incidents and their understanding of digital threats. Respondents identified various sources, including media reports (*"I have read the newspapers about cybercrimes, like hacking of information..."*) and personal experiences with cyber-attacks (*"On several occasions, I've received phone calls from people I don't know demanding money, so I really get scared"*). Some noted a lack of information within their organization, pointing to a potential communication gap (*"I think mainly most of it is self-researched or self-read because every now and then these social media platforms will talk about cyber awareness and all that"*).

Awareness was often self-acquired through research and workshops, underscoring the need for proactive learning about cyber risks (*"I just started doing a bit of research here and there. I also attended a workshop..."*).

These findings emphasize the importance of diverse sources and proactive efforts in gaining cyber threat awareness, necessary for enhancing cybersecurity strategies within organizations.

Training to Understand Cyber Risks

Training and education are key to awareness and understanding of cyber risks (Tam, et al., 2020; Stephanou & Dagada, 2008). In this sub-cluster, the focus is on the perspectives of individuals regarding their training, or lack thereof, in cyber risks. A prevalent theme among responses was the absence of formal training and awareness programmes in organizations, highlighting a need for comprehensive educational initiatives on cyber risks (*"No"; "Management, but also the employee must have a certain amount of responsibility"*).

Participants expressed the importance of promoting a culture where reporting security incidents is encouraged, indicating a proactive approach to cybersecurity (*"Sometimes, they just tell us to say if you receive a mail from a known person that you're not too sure of, don't respond to that mail"; "There is encouragement to report"*). Many have enhanced their cyber risk understanding through self-driven research and reading (*"Reading a lot, research and just becoming more and more aware with the cyber happenings"*), while others suggested workshops as an effective method for learning (*"Maybe if we are given workshops to learn about the cyber system"*).

The responses demonstrate a critical need for formal training and education on cyber risks within organizations. Encouraging incident reporting and individual efforts like reading, research, and attending workshops can contribute to building a knowledgeable and cyber-resilient workforce. These measures are important when developing a vigilant and proactive cybersecurity culture in organizations.

Cybersecurity Policies and Compliance

In cybersecurity, policies and compliance are fundamental for maintaining a secure organizational environment. Participants noted the existence of clear codes of conduct and sanctions for non-compliance, emphasizing the importance of adhering to cybersecurity policies (*"Yes, there are sanctions. We have a code of conduct"; "Some they might be given warning letters"; "There is, however, there hasn't been an instance whereby we can enforce such sanctions"; "Yes, apparently you are supposed to be charged"*).

The presence of formal cybersecurity policies varied across organizations. While some had established policies (*"Yes, we have a policy"*), most indicated a lack of such formal guidelines (*"No, it's not in our company"; "Not that I know of"*). The role of a designated cyber risk manager was considered crucial for handling cybersecurity concerns effectively (*"There's the IT manager"*), although in some instances, there was ambiguity about who was responsible (*"No, that I know of"; "Yes, it is"*).

Participants expressed the need for continuous improvement of these policies to keep pace with evolving cyber threats and to address operational limitations (*"I feel we need to do more"; "There are certain things that we are not supposed to use on these computers that we have as a company"*). Additionally, the Data Protection Officer pointed out a lack of regulation in data handling, suggesting a gap in policy enforcement and compliance (*"There has been no regulation, therefore the way the data was being handled, it really does not control"*).

Overall, these findings highlight the importance of having clear, enforceable cybersecurity policies and designated individuals for cybersecurity management. Continuous policy evaluation and enhancement are necessary to adapt to the changing cybersecurity landscape and ensure effective protection against potential risks.

5. Contributions of the Study

This study makes significant contributions in the theoretical, methodological and practical domains for cybersecurity and cyber awareness in Zambia's SME sector. Theoretically, it integrates international guidelines with Zambian realities, providing a model for adapting global cybersecurity standards to local needs. It identifies gaps in cybersecurity implementation among Zambian SMEs and proposes an adaptive compliance model within Zambia's legal framework, enhancing the understanding of cybersecurity in emerging economies.

Methodologically, the research develops Cyber Awareness and Cybersecurity guidelines specifically for Zambian SMEs, combining international best practices with local SME conditions and legal frameworks. Those guidelines, designed through literature reviews and stakeholder interviews, focus on building a cyber-aware culture and provide a structured blueprint for enhancing employee cyber awareness at all organizational levels. Such guidelines represent a significant advancement in cyber awareness research, filling a gap in the global discourse on cybersecurity methodologies.

Practically, the research advocates integrating cyber awareness into the governance and risk management of Zambian SMEs. It has generated interest among business owners and government bodies, including the Data Protection Commissioner and the Ministry of Small and Medium Enterprises, who are keen to review and potentially adopt its findings. This research offers actionable guidelines for Cyber Awareness and Cybersecurity, aligning them with business objectives and operational practices. Its adaptability across different SMEs highlights its practical utility and potential to shape national policy, positioning the research as a blueprint for elevating cyber awareness at both corporate and national levels in Zambia.

6. Limitations of the Study

The lack of current official data on the number of operational businesses in Zambia posed a foundational challenge. The study's reliance on 2012 secondary data may not accurately reflect the dynamic business sector's current state. This limitation was compounded by the study's generalized approach across the SME sector without delving into the variegated nature of different industries, each with its unique cybersecurity challenges and needs.

Another gap arises from focusing solely on digitally enabled enterprises, leaving out SMEs outside the digital sphere. This exclusion neglects a segment that, while less exposed now, remains unprotected and uninformed about cyber threats. The rapidly evolving technology and cyber threat landscape also mean the study's findings might quickly become outdated, highlighting the need for continual research and updates.

7. Ideas for Future Research

This study on cybersecurity in Zambia's SMEs opens avenues for future research, addressing its inherent limitations. Future work could explore industry-specific cybersecurity frameworks to tailor strategies for diverse sectors. Understanding the evolving nature of cyber threats through longitudinal studies will provide insights into the long-term effectiveness of cybersecurity policies.

Investigating how SMEs across varying digital integration levels navigate cyber threats is essential for a deeper understanding of national cyber resilience. Developing quantitative cybersecurity metrics and engaging in cross-country comparisons within the Southern African region could lead to more effective cybersecurity frameworks.

8. Final Word

Zambia's progression towards a connected world underscores the criticality of cybersecurity, transforming it from a mere option to an essential obligation. The rising frequency of cyberattacks and their cumulative impact necessitate proactive measures from both the government and business sectors to mitigate these risks effectively.

This research demonstrates that guidelines from international bodies like CISA and ENISA are applicable and beneficial in enhancing cyber awareness and cybersecurity knowledge within Zambia's business environment. The study's findings, derived from in-depth interviews with employees in the Zambian business sector, reveal a notable deviation from the perspectives outlined by Tischer et al. (2016). Contrary to the argument that employees predominantly assign cybersecurity responsibility to senior management, the study revealed a prevalent perception among Zambian employees that cybersecurity is a shared responsibility.

A significant revelation from the research is the lack of formal cyber awareness training within the Zambian business sector. Employees express a keen desire to learn more about cyber risks but face a void in official training resources. This gap in cyber awareness training is critical, as the knowledge it provides is fundamental for employees to safeguard their work environments effectively. The absence of initiatives from businesses and the government in conducting such training is a glaring oversight, especially considering the potential catastrophic consequences of cyber-attacks, including bankruptcy, as highlighted by several interviewees.

Another important finding is the general absence of dedicated IT personnel in Zambian businesses, a role that is essential for implementing contemporary cybersecurity measures. The prevalent reliance on antivirus software is insufficient in today's evolving threat landscape, underscoring the need for skilled IT professionals to implement more robust protections.

The impending enforcement of Zambia's Data Protection Act marks a significant advancement in the nation's legal framework. Businesses must adapt to remain compliant with the law, otherwise face potential penalties for non-compliance. The Data Protection Commissioner and the Ministry of SMEs of Zambia have shown a proactive stance, expressing interest in the outcomes of this study for potential adoption as formal guidance for

businesses. This engagement highlights the study's significant and unique contribution to Zambia's national interest and, specifically, to its business sector.

References

Akamai, (2021). State of the Internet / Security: A Year in Review, s.l.: Akamai Technologies.

Baskerville, R., Rowe, F., & Wolff, F. C. (2018). Integration of information systems and cybersecurity countermeasures: an exposure to risk perspective. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 49(1), pp 33-52.

Boyce, M. W. et al. (2011). Human performance in cybersecurity: a research agenda. In Proceedings of the Human Factors and Ergonomics Society annual meeting, 55(1), pp 1115-1119.

Chen, Y., & He, W. (2013). Security risks and protection in online learning: A survey. *The International Review of Research in Open and Distributed Learning*, 14(5).

Dagada, R. (2013). Digital banking security, risk and credibility concerns in South Africa. Second International Conference on Cybersecurity, pp 148-161.

Dagada, R. (2014). Legal and policy aspects to consider when providing information security in the corporate environment. Thesis submitted in accordance with the requirements for the degree of Doctor of Philosophy in Information Systems at the University of South Africa, November 2014.

Dagada, R. (2021). *Digital Commence governance in the era of fourth industrial revolution in South Africa*. Pretoria: Unisa Press.

Federal Bureau of Investigation. (2020). IC3 Annual Report, s.l.: Federal Bureau of Investigation.

Government of Zambia. (2021). The Cyber Security and Cyber Crimes Act No. 2 of 2021, Lusaka, Zambia: Government of Zambia.

Government of Zambia. (2021). The Data Protection Act No. 3 of 2021, Lusaka, Zambia: Government of Zambia.

Gundu, T. (2019). Acknowledging and reducing the knowing and doing gap in employee cybersecurity compliance. 14th International Conference on Cyber Warfare and Security, pp94-102.

Ifinedo, P. (2023). Effects of security knowledge, self-control, and countermeasures on cybersecurity behaviors. *Journal of Computer Information Systems*, 63(2), pp 80-396.

Kiger, M. E., & Varpio, L. (2020). Thematic analysis of qualitative data: AMEE Guide, No. 131. *Medical Teacher,* 42(8), pp 846-854.

Kortjan, N., & von Solms, R. (2014). A conceptual framework for cyber-security awareness and education in South Africa. *South African Computer Journal*, 52(1), pp 29-41.

Kozak, S. (2017). The role and importance of the small business sector in the economic development of the Mazowieckie Province. *Scientific Journals of the University of Natural Sciences and Humanities, Series Administration and Management*, 41(114), pp 61-70.

Morgan, S. (2022). *Official Cybercrime Report*. Northport: Cybersecurity Ventures.

National Assembly of Zambia (2022). Information Brief on Cyber Security and Cybercrime Trends in Zambia: Research Department. Lusaka. National Assembly of Zambia.

Nifakos, S. et al. (2021). Influence of human factors on cybersecurity within healthcare organisations: A systematic review. *Sensors*, 21(15), pp 5119.

Ponemon Institute. (2020). *Cost of a Data Breach Report 2020*. New York: IBM Security.

Rajasekharaiah, K. M., Dule, C. S., & Sudarshan, E. (2020). Cyber Security Challenges and its Emerging Trends on Latest Technologies. *IOP Conference Series: Materials Science and Engineering*, 2(981), p 022062.

Schatz, D., Bashroush, R., & Wall, J. (2017). Towards a More Representative Definition of Cyber Security. *Journal of Digital Forensics, Security and Law*, 12(2), p 8.

Shafqat, N., & Masood, A. (2016). Comparative Analysis of Various National Cyber Security Strategies. *International Journal of Computer Science and Information Security*, 14(1), p 129.

Shaw, R. S., Chen, C. C., Harris, A. L., & Huang, H.-J. (2009). The impact of information richness on information security awareness training effectiveness. *Computers & Education*, 52(1), pp 92-100.

Slusky, L. (2020). Cybersecurity of online proctoring systems. *Journal of International Technology and Information Management*, 29(1), pp 56-83.

Stephanou, T. & Dagada, R. (2008). The impact of information security awareness training on information security behaviour: the case of further research. *ISSA 2008*, University of Johannesburg, 2 to 4 July 2008.

Tam, K., Moara-Nkwe, K., & Jones, K. (2020). The use of cyber ranges in the maritime context: Assessing maritime-cyber risks, raising awareness, and providing training.. s.l.:s.n.

World Economic Forum. (2018). Cybersecurity: The $1 Trillion Opportunity, s.l.: World Economic Forum.

Yildirim, E. (2016). The importance of information security awareness for the success of business enterprises.

In Advances in Human Factors in Cybersecurity. *Proceedings of the AHFE 2016 International Conference on Human Factors in Cybersecurity*, July 27-31, 2016, Walt Disney World®, Florida, USA, pp 211-222.

Zambia Development Agency. (2020). Promoting SME competitiveness in Zambia. Lusaka, Zambia: Zambia Development Agency.

ZICTA. (2022). Collaborative Framework: For the Oversight of Digital Financial Services in Zambia, Lusaka, Zambia: Zambia Information and Communications Technology Authority (ZICTA), Bank of Zambia (BoZ) and the Rural Finance Expansion Programme (RUFEP).

Coming Back Down to Earth: A Grounded Look at Space Cybersecurity in Southeast Asia

Shantanu Sharma
Center of Excellence for National Security, S. Rajaratnam School of International Studies, Singapore

shantanu.sharma@ntu.edu.sg

Abstract: This paper aims to open discussion regarding the absence of cybersecurity efforts for space systems in Southeast Asia and its implications. Although there is a consensus on the need for action, Southeast Asian states have not taken any concrete steps to address the challenge of space cybersecurity. Lacking sovereign capabilities, regulatory structures, and initiatives to tackle complex accidents that are waiting to happen, Southeast Asia is not adequately equipped to deal with the growing risk of attack on its space ecosystem. With increasing militarization of space programs, exponentially growing dominance of private sector, and reduced barriers of entry for malicious non-state actors, delaying any action for strengthening space cybersecurity can produce detrimental impact given the necessity of space sector for civil, commercial, and national sustenance in the region. Southeast Asia is in a unique position where the nascency of space sector and lack of rigid structures allows flexibility to learn, translate, and adapt existing initiatives and frameworks. To be on a path to bridge the gap between reality and the ideal, the region should focus on articulating principles and guidelines. The author proposes realistic policy recommendations and directions Southeast Asian states and the region at-large should pursue for promoting medium and long-term sustainability and usability, and for maintaining cybersecurity of space assets.

Keywords: Space, Cybersecurity, Southeast Asia, Cyber threat, Satellite

1. Introduction: "New Space" and New Opportunities

A renaissance in the space industry over the past few decades, driven by technology innovation, private sector investment, entrepreneurial activities, and diminished launch cost, has resulted in affordability and "democratization of space" (Wesler IV 2016). Outer space has become accessible not just to large enterprises and states but also to small state and non-state operators including start-ups and universities.

Technological miniaturization brought about by advancement in communication technologies and microelectronics have led to development of complex, smaller, and power efficient satellites. Furthermore, commercialization and global supply chains have produced an "assembly line" model for space system development. This has resulted in better access to foreign technologies (Upadhyay 2023), modern project management, agile production of modular commercial-off-the-shelf (COTS) components, payload specialization, shorter development process, and introduction of "as-a-Service" models.

Given its complex geography, space sector is a necessity for Southeast Asian states for civil, commercial, and national sustenance. States in the region use space systems to provide:

- earth observation for minerals, fisheries (The Economist 2023) and seasonal agricultural monitoring, disaster mitigation and management (Cowing 2023), and weather forecasting
- remote sensing and navigation for aviation and maritime sector
- telecommunications, internet (Raj 2023), and broadcasting for inaccessible islands in the archipelago.

Confronted with novel challenges like climate change, rising tensions in Taiwan Strait, increasing disputes in South China Sea, sudden shifts in supply chains (Tan 2023) due to friend-shoring and nearshoring, and growing digital population; interest in space sector have intensified in the region (Sarma 2019).

The current state of space ecosystem in Southeast Asia is fraught with fragmented capabilities and national objectives, and unequal legal and regulatory commitments (See Figure 1). Some states in Southeast Asia have established institutions and programmes, others are in the earliest stages of structuring their own. While no Southeast Asian state has an independent launch capability, five have national space programs – Indonesia, Malaysia, the Philippines, Thailand, and Vietnam. However, Brunei, Cambodia, Laos, and Myanmar have limited or no activities in space (Verspieren 2023). Whereas Singapore, geared towards a socio-economic angle, is active in developing its space sector despite not having a national agency.

Around the globe space systems have become deeply integrated with terrestrial digital infrastructure for global economic, militaristic, societal, and governmental activities. For many industries, space systems represent single point of failure (Falco 2018) with increasing relevance in countries' critical infrastructure (Carlo 2021).

Although Southeast Asia's space sector is still relatively nascent, threats to its space assets are no longer speculative. In June 2018, a campaign launched from China targeted companies in Southeast Asia and Unites States, infected satellite operators, defense contractors, and telecommunications companies with the aim to disrupt data traffic and change position of orbiting satellites (Menn 2018). The attack mounted did not cease until a year later. Indonesia, Malaysia, the Philippines, and Vietnam are just the latest victims of increasing attacks on space infrastructure in Southeast Asia (Symantec 2019).

Current literature focusing on space cybersecurity at-large do not consider nascency of space sector, fragmented capabilities, and disjointed objectives of states in Southeast Asia. Although, Verspieren (2022, 2023), Sarma (2019), Takahashi and Funabiki (2022), and Jones (2014) examine regional capabilities and structures and provide recommendations for encouraging growth of space ecosystem, the papers do not tackle security of space assets in the region. This paper therefore tries to fill the gap in current literature and aims to provide grounded policy recommendations that states in Southeast Asian should pursue to confront the growing cyber threat to space systems by leveraging their unique position and existing regional structures. The paper is structured as follows - after providing a brief overview of space ecosystem in Southeast Asia (Section 1), the paper details the importance of space system especially for military and national security (Section 2), the paper then focuses on challenges for implementing cybersecurity measures for space systems (Section 3 and 4); and lastly, a realistic plan for developing space cybersecurity measures in Southeast Asia is explored (Section 5).

Table 1: Comparison between major space powers in Southeast Asia[1]

	Singapore	Indonesia	Malaysia	The Philippines	Thailand	Vietnam
National Space Agency (Year of establishment)	None	BRIN (since 2022) Previously: LAPAN (1963)	MYSA (2019) Previously: ANGKASA (2002)	PhilSA (2019)	None. Consortium of 3 space related institutions including GISTDA (2000)	None. 2 Space-related institutes – STI (2006) and VNSC (2011) – under Vietnam Academy of Science and Technology
National Space Council (Years active)	None	None. Previously: DEPANRI (1963-2014)	None	Philippines Space Council (2019- Present)	None	None
National Space Law	None	Law no. 21 of 2013 on Space Activities	None	Philippines Space Act (2019)	None	None
National Space Policy or Strategy (Latest)	None	Presidential Regulation no. 45 of 2017 concerning the Master Plan for Space Activities (2016-2040)	National Space Policy 2030 (2017)	Philippines Space Development and Utilization Policy (2019)	National Space Master Plan 2017-2036	Strategy of Space Science and Technology Development and Application to 2030 (2021)
UN Space Treaties Ratified and Signed	Ratified: OST, ARRA, LIAB Signed: REG	Ratified: OST, ARRA, LIAB, REG	Signed: OST, ARRA	Ratified: MOON Signed: OST, ARRA, LIAB	Ratified: OST, ARRA	Ratified: OST Signed: ARRA
Government Space Budget (2020, in million USD)	35	303	22	57	94	21
Independent Space Launch Capabilities	No	No	No	No	No	No
Launch Year of 1st Satellite At Least Partially Made by Nationals	2011 (X-SAT)	2007 (LAPAN-TUBSAT)	2000 (TiungSAT-1)	2016 (DIWATA-1)	1998 (THAI-PAHT)	2013 (PicoDragon)
Launch Year of 1st Domestically Produced Satellite	2011 (X-SAT)	2015 (LAPAN-A2)	2023 (A-SEANSAT-PG 1)	None	None	2013 (PicoDragon)

2. Militarization of Space: Space for Defense and Defense for Space

Lower barrier of access to space has led to increasing number of satellites with multiplication and diversification of actors and their objectives have resulted in unprecedented proliferation of outer space, leading to congestion of LEO and near saturation of the frequency spectrum (De Neve 2022).

Space geopolitics mirrors Earth – there is a multipolar international order, intensifying tensions between competing powers, and the presence of commercial actors disrupting and increasingly dominating the space sector. Given the importance of space systems for sustenance, threat of degradation, disruption, and destruction by malign actors has become a national security issue and outerspace has become congested, contested, and competitive. To secure themselves, major powers are shoring up arsenal of counterspace weapons as a form of deterrence, leading to greater contestation and militarization of outer space.

A shift to multi-domain warfare strategies has increased the requirement for support from space-based platform for C4ISR (Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance) (Verma 2023) capabilities and to obtain operational superiority against A2/AD (Anti-access/area denial) capabilities. This reliance on space systems has materialized as "space for defense" – with NATO declaring space

[1] Derived from (Verspieren 2022), (UNOOSA 2023), and (Ignatius 2023). Abbreviations: *BRIN* - National Research and Innovation Agency, *LAPAN* – National Institute of Aeronautics and Space, *MYSA* – Malaysia Space Agency, *ANGKASA* – National Space Agency, *PhilSA* – Philippines Space Agency, *GISTDA* – Geo-Informatics and Space Technology Development Agency, *STI* – Space Technology Institute, *VNSC* – Vietnam National Space Centre, *OST* - Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer Space, including the Moon and Other Celestial Bodies (Outer Space Treaty 1967), *ARRA* - Agreement on the Rescue of Astronauts, the Return of Astronauts and the Return of Objects Launched into Outer Space (Rescue Agreement 1968), *LIAB* - Convention on International Liability for Damage Caused by Space Objects (Liability Convention 1972), *REG* - Convention on Registration of Objects Launched into Outer Space (Registration Convention 1975), and *MOON* - Agreement Governing the Activities of States on the Moon and Other Celestial Bodies (Moon Agreement 1979).

as the fifth operational domain (NATO 2023), and EU identifying space as a strategic domain (European Commission 2023).

Increasing operational dependence on space systems, both commercially and military owned, marks them as a target for an adversary trying to gain asymmetrical advantage. The issue was brought to the forefront during the Russian invasion of Ukraine. On 24 February 2022, a cyberattack on Viasat's KA-SAT satellite communication service was conducted by Russia, an hour before the ground invasion to cripple Ukrainian Army communications which relied on the satellite infrastructure. This also resulted in unanticipated consequences, affecting broadband services to tens of thousands of users in France, Hungary, Greece, Italy, Poland, and Germany, and affecting Enercon's remote monitoring and control access of 5,800 wind turbines (Poirier 2022).

Reliance on space systems also requires "defense for space". Following the United States, European states have reacted to threats in outerspace through militarization. Germany, Italy, and UK have established space commands (Rome 2021), and Spain and France (Machi 2022) have established air and space force, with more states to follow suit. US also introduced bipartisan Space Infrastructure Act to designate space sector to be the seventeenth critical infrastructure (Fortinsky 2023). A similar reaction, initiation of militarization of space program, is playing out in space-faring nations surrounding Southeast Asia – India (Kanwar et al. 2023), China (Stokes et al 2020), South Korea (Byun et al 2023), Japan (Kyodo News 2023), and Australia (Davis 2023).

3. Cyber and Space Infrastructure

Cyber counterspace activities is not foreign to Southeast Asia, despite states having minimal space assets and no active mega-constellations. Notably, in 1986, Indonesia was the first state accused of satellite eavesdropping violation (Pavur and Martinovic 2022); and in 1996, Indonesia was the first state to deliberately use a satellite for signal jamming another satellite (Fergusson and Wong 2010).

Due to extreme and unexpected environmental conditions in space and interference arising from ground-to-satellite communication (Pultarova 2023), it is challenging to distinguish unintentional failure or hardware malfunction from electronic and cyber-attacks (Wheeler et al. 2018). This results in lack of direct attribution to an attacker, providing political maneuverability for the attacker which can help evade any direct retaliation, escalation, or legal and political implications.

Challenges in attribution, possibility for deniability, controlled escalation, low operational cost and ease of access, ability to create reversible effects, and assurance of non-destructive attack below the threshold of armed conflict (Bingen, Johnson and Young 2023) make cyber-attacks on space systems stand out as the ideal option to create significant disruptions and instill damaging effect on military, commercial, and civilian operations.

Ambiguities in international cyber policies, regulations, and standards, exacerbated by shortcomings and creation of harmful technical knowledge gaps in adaptation of guidelines from terrestrial cyber assets to space cyber assets; creates an ideal space for adversaries to operate in and exploit (Pavur and Martinovic 2022). To date, almost 30 countries, including many non-spacefaring nations, have demonstrated some degree of cyber-offensive counterspace capabilities (Pavur and Martinovic 2022).

4. The Challenge of Space Cybersecurity

Increased digitalization through advanced on-board processing, software defined radios (SDRs), packet-based protocol, and cloud enabled high-performance computing has led to rapidly expanding attack surface available for cyberattacks on space systems. This has resulted in a significant spike in attacks like eavesdropping (DEFCON Conference 2020), malware injection (Francis 2022), and jamming (Laursen 2023), and also increased significant risk of backdoor, denial of service, and hijacking attacks (Hadley 2023). Even cyberattacks instigated by non-state actors and even independent hackers have risen in frequency, complexity, and magnitude (Rajagopalan 2019, Pavur and Martinovic 2022).

As satellites operate in challenging environments with different mission objectives, they require bespoke hardware to operate. Combined with proliferation of COTS components, a unique situation is created where vulnerabilities likely apply to many platforms but applying patches requires tailored modifications (Pavur 2021).

The loss of physical access to space assets after launch also limits the ability to perform upgrades, repairs, or audits. This translates to more acute security/performance trade-offs to be considered for space systems as compared to terrestrial systems. Satellites are highly constrained devices with limited power, fuel, and compute capabilities. Adopting stronger cybersecurity principles such as stronger encryption would incur bandwidth, storage, and compute cost, which would directly compete with core functionalities (Pavur and Martinovic 2022).

Since the value chain has evolved into a shared ownership model, security of space data and systems is a shared responsibility between satellite operators, carriers, and satellite manufacturers (Grady et al. 2022). With the blurring of boundaries due to inter-dependent multi-stakeholder collaboration, it is becoming increasingly difficult to clearly assign security ownership and impose regulations.

Taking note of the stacking complexities and vulnerabilities in space systems, militaries have also started building up cyber counterspace capabilities. For example, a CIA document leaked on Discord revealed Chinese build-up of sophisticated cyber weapons to target satellites which would allow China "to seize control of a satellite, rendering it ineffective to support communications, weapons, or intelligence, surveillance, and reconnaissance systems" (Srivastava, Schwartz and Sevastopulo 2023).

5. Next Steps for Southeast Asia: Leveraging Existing Structures

Southeast Asia currently finds itself in a unique position where the nascency of space sector and lack of rigid structures allows flexibility to learn, translate and adapt existing initiatives and frameworks put forth by more advanced space powers like USA, UK, Germany, China, India, Japan, and France. Exploiting this limited window of opportunity in a meaningful way would require collaboration, extending partnerships, strengthening existing initiatives, and aiming for regulations.

Walking the road towards concrete and clearly defined regulations require the first steps taken to be guidelines, norms, and principles, which are currently absent in Southeast Asia. It is vital to enshrine multistakeholder process at the heart of the norm development process. This requires engaging with the industry, international partners, and regional governments.

With international bodies like the UN OEWG paralyzed due to geopolitics, countries are leaning towards minilateral arrangements. Regional-level forums to discuss space-related issue like Japan-led APRSAF (Asia-Pacific Regional Space Agency Forum), China-led APSCO (Asia-Pacific Space Cooperation Organisation), Spatial Information Corridor under China's Belt and Road Initiative, and norm-based US-led Artemis Accords do exist, but they are subject to the competitive geopolitical dynamics at play globally. Despite a handful of countries that are members, all Southeast Asian states are not fully represented in any forum. There is no mechanism for these different forums focused on Southeast Asia to consult, cooperate, or communicate with each other.

As existing Southeast Asian structures are not concrete enough to collaborate, states are relying on cooperation of international partners like China, Japan, India, and the US to help build space capabilities. For example, in March 2023, United States, Japan, Malaysia, and Thailand conducted multilateral military exercise responding to simulated space attack (Inoue and Shiga 2023). As the space sector is still nascent in Southeast Asia, it would be important to focus on low hanging fruits for capacity and knowledge building and raising awareness, like constructing win-win collaborations through peaceful uses of outerspace like humanitarian assistance and disaster relief (HADR) objectives (U.S. Department of State 2023) and creating data centres in the region to process and analyse earth observational data.

To promote medium and long-term sustainability, usability, and security of space, existing Southeast Asian structures need to be strengthened and collective action is required. ACICE (ASEAN Cybersecurity and Information Centre of Excellence) and SCOSA (ASEAN Sub-Committee on Space Technology and Applications) should be utilized to discuss and put-forth guidelines for space cybersecurity and translate existing international frameworks for the region. For confidence-building measures, communication, cooperation, and collaboration between APSCO and APRSAF should be encouraged. More regional coordination and centres for collaboration are needed. The collective, coordinated, and unified voice can be more impactful, especially in international processes like UN OEWG and Space-ISAC.

Provided below are some concrete recommendations for Southeast Asian states. These should serve as the foundation for creating guidelines, norms, and principles to strengthen the cybersecurity of space assets -

5.1 Fostering Talent

Space is an inherently cross-disciplinary field. Even established space powers, like Japan, with advanced capabilities and significant resource allocation, are facing talent shortage to meet the growing demand (Verspieren 2021). The unmet demand will result in Southeast Asian states cannibalizing talent off of each other which would dampen the space ecosystem. Cross-disciplinary training and objective-based exchanges are needed to sustainably foster talent, which can be achieved in the region through replicating examples of existing memorandum of understanding (Space Foundation 2022), cross-university efforts, international research

initiatives, and collaboration for peaceful operations (JICA 2021) like climate monitoring and HADR (Garekar 2023).

5.2 Enforcing Basic Cybersecurity Principles and Practicing Cyber Hygiene

Willbold et al. (2023) highlighted the abysmal state of space cybersecurity. They noted that basic cybersecurity principles are missing, and in practice operators were instead still relying on outdated 'security through obscurity'. They also highlighted lack of authentication protocols, broadcasting of signal without encryption, lack of penetration testing, and missing separation of code and data.

To cultivate cyber hygiene, organizational security culture, and to raise awareness, Southeast Asia states can follow the examples of NIST IR 8270, 8401, 8441 (Scholl and Suloway 2021, Lightman, Suloway, and Brule 2022, McCarthy et al. 2023), and especially of UK Space Cybersecurity Toolkit which provides information on potential cyber threats, cybersecurity standards to adopt, list of relevant authorities, reporting obligations and process, and outlines impact assessment methodologies (UK Space Agency 2020). States should also promote broad set of frameworks and basic security principles like Zero Trust Architecture framework (Easley 2023), DevSecOps (Leonard 2021), and end-to-end encryption (E2E) to reduce the attack surface.

5.3 Fixing Supply Chain Vulnerabilities

The supply chain has shifted towards open-source software and COTS hardware for faster, cheaper, scalable, and standardized space system manufacturing and deployment. This can lead space systems vulnerable to attacks like backdoors and code injection. Attack on supplier side is not uncommon. Maximum Industries, a SpaceX supplier, was hit with a ransomware attack on March 2023 (Greig 2023).

Since supply chain have become global, ideally countries should consider extending cybersecurity certification and labelling scheme to space hardware and software. Nevertheless, space ecosystem in Southeast Asia is still in early stages of development, imposing compliance on local small business and start-ups could create additional cost, and in-turn create an unfair advantage for established international players (Strohmier et al. 2022). However, it is important for states to list and impose vendor check and certification requirement for space systems being used in military or dual-use capacity. US CMMC 2.0 certification (U.S. Department of Defense 2021) requirement is an example that Southeast Asian states can replicate, especially for satellites required to operate in D3SOE (Denied, Degraded, and Disrupted Space Operational Environment) capacity. Thailand, which is considering using space technology for defense purposes, is already contemplating legislation to curb foreign influence in COTS production (Inoue and Shiga 2023).

5.4 Modernizing Legacy Hardware

Lack of physical access to space-borne assets creates a unique situation where it is impossible to maintain aging hardware. These legacy assets can prove as an easy entry point for attackers (Kallberg 2018). States in Southeast Asia should list and modernise their aging space systems, including ground systems, when possible (Erwin 2021).

5.5 National Security and the Private Sector

The role of commercially-owned satellites in critical government operations should prompt countries to start thinking about guidelines and regulations that prioritize security. After Russia's cyberattack on Ukrainian space infrastructure, Ukraine replaced many communication links using SpaceX's Starlink constellation for communication (Weeden and Samson 2023), intelligence gathering, and coordinating drone strikes (Satariano et al. 2023). In 2022, Elon Musk after conversation with Russian officials (Davis 2023), deactivated Starlink satellites near Crimea to prevent a Ukrainian offensive operation (Kim 2023), leading to loss of lives (Farrow 2023). The private sector has become increasingly important and disruptive, prompting national security concerns, and provoking a wide range of reactions. It is noted that 80% of US military communications rely on commercial satellites (Becht and Trompille 2019). Concerns about Starlink dominance has influenced the EU to launch its own 2.4 Billion Euros satellite constellation for civilian and military use (Kim 2023). Whereas countries like Japan (Umeda 2023) and United States (Schogol 2023) are exploring options to use SpaceX products for militaristic purposes.

The states in Southeast Asia can look towards 8 (EUR-Lex 2023) and Critical Entities Resilience Directive (European Parliament and The Council of the European Union 2022) as foundation for coming up with broader principles.

5.6 Creating Cybersecurity Testbeds

As there are limited number of satellite development centres in Southeast Asia, the region needs to develop replicable research testbeds for satellite platform exploitation. Testbeds allows for controlled and repeatable experiment, improves access to technology for researchers who do not own and operate satellites (Pavur and Martinovic 2022), and combined hands-on experience with educational courses allow for better training for engineers. Similar bug bounty and penetration testing program should be explored like those initiated by SpaceX (Duffy 2022); European Space Agency's CYSAT 2023, where researchers took control of ESA nanosatellite (Thales Group 2023); and DEF CON 2023 Hack-a-Sat (Antoniuk 2023), where a satellite was launched for first ever capture-the-flag competition on a space-borne asset (Vasquez 2023).

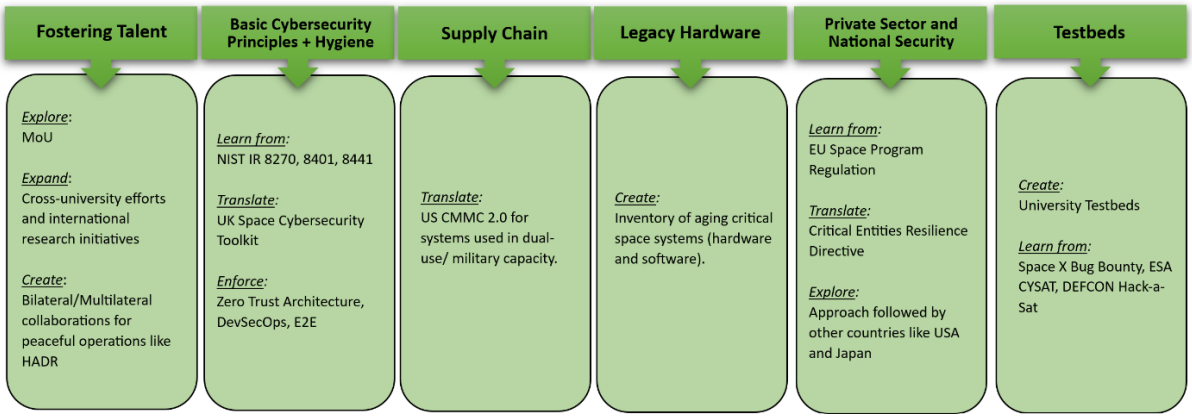


Figure 1: Summary of recommendations and directions Southeast Asian states should pursue

6. Conclusion

Around the globe space systems have become deeply integrated with terrestrial digital infrastructure for global economic, militaristic, societal, and governmental activities. The cyberattack on Viasat satellite brough the issue of cybersecurity of space assets to the forefront. States in Southeast Asia are not adequately equipped to deal with the growing risk of attack on their space ecosystem. However, Southeast Asia currently finds itself in a unique position where the nascency of space sector and lack of rigid structures allows flexibility to learn, translate and adapt existing initiatives and frameworks. Exploiting this limited window of opportunity in a meaningful way would require collaboration, extending partnerships, strengthening existing initiatives, and aiming for regulations. The paper provides some recommendations Southeast Asian states should pursue (See Figure 1) for increasing stability, strategic autonomy, building sovereign capabilities, and securing its space assets from cyberattacks. The recommendations provided should serve as foundation for creating guidelines, norms, and principles. Further research and dialogue are needed to create regulations and standards for cybersecurity of space assets.

Acknowledgement

The author would like to express his gratitude to Manoj Harjani (Research Fellow, IDSS, RSIS) for providing invaluable feedback and comments.

References

Antoniuk, D. (2023) "This new satellite enters orbit with one mission: To get abused by hackers", *Recorded Future*, 7 June. Available at: https://therecord.media/new-satellite-enters-orbit-to-get-hacked (Accessed at: 17 November 2023).

Becht, O. and Trompille, S. (2019) "Information report n°1574", *French Commission de la Défense Nationale et des Forces Armées*, pp. 70-71. Available at: https://www.assemblee-nationale.fr/dyn/opendata/RINFANR5L15B1574.html (Accessed at: 17 November 2023).

Bingen, K., Johnson, K. and Young, M. (2023). "Space Threat Assessment 2023", *CSIS*, pp. 5-7. Available at: https://csis-website-prod.s3.amazonaws.com/s3fs-public/2023-04/230414_Bingen_Space_Assessment.pdf (Accessed at: 17 November 2023).

Byun, C., Ahn, T., Choi, S. and Handberg, R. (2023) "Developing the Direction of Military Space Capabilities in South Korea." *Journal of Indo-Pacific Affairs*, *6*(4), pp. 106-112.

Carlo, A. (2021) 'Cyber threats to space communications: space and cyberspace policies', *Outer Space and Cyber Space: Similarities, Interrelations and Legal Perspectives*, pp.55-66.

Cowing, K. (2023) 'Satellite-based Disaster Early Warning Systems Can Improve Evacuation Measures In Remote Asian Communities', *SpaceRef*, 5 July. Available at: https://spaceref.com/newspace-and-tech/satellite-based-disaster-early-warning-systems-can-improve-evacuation-measures-in-remote-asian-communities/ (Accessed at: 15 November 2023).

Davis, C. R. (2023) "Elon Musk blocked Ukraine's access to Starlink near Crimea after speaking with Russian officials, biographer says", *Business Insider*, 8 September. Available at: https://www.businessinsider.com/elon-musk-blocked-ukraine-starlink-access-crimea-russia-war-putin-2023-9 (Accessed at: 17 November 2023).

Davis, M. (2023) "A higher place for space in the defence strategic review", *Australian Strategic Policy Institute The Strategist*, 8 May. Available at: https://www.aspistrategist.org.au/a-higher-place-for-space-and-the-defence-strategic-review/ (Accessed at: 16 November 2023).

De Neve, A. (2022) "Crowded and Dangerous Orbits: European Space Governance at a Time of Potentially Saturating Programs", *Notes de l'Ifri*, pp. 19-20. Available at: https://www.ifri.org/sites/default/files/atoms/files/a._de_neve_crowded_dangerous_orbits_feb_2022.pdf (Accessed at: 15 November 2023).

DEFCON Conference (2020) *DEF CON Safe Mode - James Pavur - Whispers Among the Stars.* Available at: https://www.youtube.com/watch?v=ku0Q_Wey4K0 (Accessed at: 16 November 2023).

Duffy, K. (2022) "SpaceX says researchers are welcome to hack Starlink and can be paid up to $25,000 for finding bugs in the network", *Business Insider*, 15 August. Available at: https://www.businessinsider.com/spacex-starlink-pay-researchers-hack-bugs-satellite-elon-musk-2022-8 (Accessed at: 17 November 2023).

Easley, M. (2023) "Space Force awards contract to protect space systems, data with zero-trust approach", *Defense Scoop*, 3 October. Available at: https://defensescoop.com/2023/10/03/space-force-xage-security-zero-trust/ (Accessed at: 17 November 2023).

Erwin, S. (2021) "Space Force grappling with aging infrastructure used to operate satellites", *Space News*, 19 September. Available at: https://spacenews.com/119100-2/ (Accessed at: 17 November 2023).

EUR-Lex (2023) *EU space programme (2021–2027) – European Union Agency for the Space Programme.* Available at: https://eur-lex.europa.eu/EN/legal-content/summary/eu-space-programme-2021-2027-european-union-agency-for-the-space-programme.html (Accessed at: 17 November 2023).

European Commission (2023) "EU Space Strategy for Security and Defence" Available at: https://defence-industry-space.ec.europa.eu/eu-space-policy/eu-space-strategy-security-and-defence_en (Accessed at: 15 November 2023).

European Parliament and The Council of the European Union (2022) DIRECTIVE (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the Resilience of Critical Entities and Repealing Council Directive 2008/114/EC. *Off. J. Eur. Union*, *333*, p.164. ELI: http://data.europa.eu/eli/dir/2022/2557/oj

Falco, G. (2018) 'The vacuum of space cyber security', *AIAA SPACE and Astronautics Forum and Exposition*, p. 5278.

Farrow, R. (2023) "Elon Musk's Shadow Rule", *The New Yorker*, 21 August. Available at: https://www.newyorker.com/magazine/2023/08/28/elon-musks-shadow-rule (Accessed at: 17 November 2023).

Fergusson, J. and Wong, W.W. eds. (2010). *Military space power: a guide to the issues*. Bloomsbury Publishing USA.

Fortinsky, S. (2023) "Bipartisan bill designates space as critical infrastructure sector", *The Hill*, 27 July. Available at: https://thehill.com/homenews/space/4123413-bipartisan-bill-designates-space-as-critical-infrastructure-sector/ (Accessed at: 16 November 2023).

Francis, J. (2022) "Briefing 3: Wiper Malware: An Increasing Threat to Satellite and IoT Enabled Capabilities", *Kratos Defense Threat Briefing*, 12 December. Available at: https://www.kratosdefense.com/constellations/articles/threat-briefing-3-wiper-malware-an-increasing-threat-to-satellite-and-iot-enabled-capabilities (Accessed at: 16 November 2023).

Garekar, B. (2023) "Haze detection, aviation and maritime safety discussed in first S'pore-US space dialogue", *The Straits Times*, 13 October. Available at: https://www.straitstimes.com/world/ways-to-detect-haze-enhance-aviation-maritime-safety-discussed-in-first-s-pore-us-space-dialogue (Accessed at: 17 November 2023).

Grady, B., Camp, C. V., Muruganandham, S., and Placido, C. (2022) "Space Cybersecurity - Current State and Future Needs", *Northern Sky Research,* April, pp. 4-11. Available at: https://www.nsr.com/wp-content/uploads/2022/04/NSR-Space-Cybersecurity-White-Paper-FINAL.pdf (Accessed at: 17 November 2023).

Greig, J. (2023) "FBI, Air Force warn of cyberattacks on space industry by 'foreign intelligence operations'", *Recorded Future*, 19 August. Available at: https://therecord.media/fbi-warns-of-space-cyberattacks (Accessed at: 17 November 2023).

Hadley, G, (2023) "'Backdoor' to Attack Satellites: CSO Sees Cyber Risks in Space Force Ground Systems", *Air & Space Forces Magazine*, 31 January. Available at: https://www.airandspaceforces.com/backdoor-to-attack-satellites-cso-highlights-ground-networks/ (Accessed at: 17 November 2023).

Ignatius, C. (2023) "ANGKASA-X Launches First Homegrown Satellite", *BusinessToday Malaysia*, 28 June. Available at: (Accessed at: 20 January 2024)

Inoue, K., Shiga, Y. (2023) "Southeast Asia's space race chases wins in tourism, communications", *Nikkei Asia*, 11 May. Available at: https://asia.nikkei.com/Business/Aerospace-Defense-Industries/Southeast-Asia-s-space-race-chases-wins-in-tourism-communications (Accessed at: 17 November 2023).

JICA (2021) "Collaboration with JAXA to develop human resources in "Space Technology Utilization for SDGs" in Southeast Asia", 8 October. Available at: https://www.jica.go.jp/Resource/english/news/field/2021/20211008_01.html (Accessed at: 17 November 2023).

Jones, Z.P. (2014) *Southeast Asian space programs: motives, cooperation, and competition* (Doctoral dissertation, Monterey, California: Naval Postgraduate School). Available at: https://apps.dtic.mil/sti/tr/pdf/ADA619544.pdf (Accessed at: 20 January 2024).

Kallberg, J. (2018) "Why older satellites present a cyber risk", *C4ISRNET*, 29 December. Available at: https://www.c4isrnet.com/opinion/2018/12/28/why-older-satellites-present-a-cyber-risk/ (Accessed at: 17 November 2023).

Kanwar, V., Kapil, S., Suryavanshi, H., Reddy, K., and Chatterjee, C. (2022) "Space for defence in India", *PricewaterhouseCoopers and SatCom Industry Association India*, October, pp. 19-26. Available at: https://www.pwc.in/assets/pdfs/aero_defence/space-for-defence-in-india.pdf (Accessed at: 16 November 2023).

Kim, V. (2023) "Elon Musk Acknowledges Withholding Satellite Service to Thwart Ukrainian Attack", *The New York Times*, 8 September. Available at: https://www.nytimes.com/2023/09/08/world/europe/elon-musk-starlink-ukraine.html (Accessed at: 17 November 2023).

Kyodo News (2023) "Japan adopts space security policy, vows to expand defense use", 13 June. Available at: https://english.kyodonews.net/news/2023/06/caac42b9ada8-japan-adopts-space-security-policy-vows-to-expand-defense-use.html (Accessed at: 16 November 2023).

Laursen, L. (2023) "Satellite Signal Jamming Reaches New Lows", *IEEE Spectrum*, 18 May. Available at: https://spectrum.ieee.org/satellite-jamming (Accessed at: 16 November 2023).

Leonard, J. (2021) "DevSecOps in space: the challenges of updating satellites on-orbit", *Computing*, 10 June. Available at: https://www.computing.co.uk/analysis/4032657/devsecops-space-challenges-updating-satellites-orbit (Accessed at: 17 November 2023).

Lightman, S., Suloway, T., and Brule, J. (2022) "Satellite Ground Segment", *National Institute of Standards and Technology Interagency Report*, NIST IR 8401. Available at: https://doi.org/10.6028/NIST.IR.8401 (Accessed at: 17 November 2023).

Machi, V. (2022) "France puts space at top of national — and European — security priorities", *Defense News*, 14 March. Available at: https://www.defensenews.com/space/2022/03/14/france-puts-space-at-top-of-national-and-european-security-priorities/ (Accessed at: 14 November 2023).

McCarthy, J., Mamula, D., Brule, J., Meldorf, K., Jennings, R., Wiltberger, J., Thorpe, C., Dombrowski, J., Lattin, O.R. and Sepassi, S. (2023) "Cybersecurity Framework Profile for Hybrid Satellite Networks (HSN)", *National Institute of Standards and Technology, NIST Interagency or Internal Report (IR) NIST IR*, *8441*(2023), p.28. DOI: 10.6028/NIST.IR.8441.

Meirizal, A. and Putri, D. J. (2022) "The Challenge of ASEAN Institutionalism in Outer Space", *ASEAN Studies Center Universitas Gadjah Mada*, 5 October. Available at: https://asc.fisipol.ugm.ac.id/2022/10/05/the-challenge-of-asean-institutionalism-in-outer-space/ (Accessed at: 20 January 2024).

Menn, J. (2018) 'China-based campaign breached satellite, defence companies – Symantec', *Reuters*, 20 June. Available at: https://www.reuters.com/article/us-china-usa-cyber/china-based-campaign-breached-satellite-defense-companies-symantec-idUSKBN1JF2X0 (Accessed at: 14 November 2023).

NATO (2023) "NATO's approach to space", 23 May. Available at: https://www.nato.int/cps/en/natohq/topics_175419.htm (Accessed at: 15 November 2023).

Pavur, J. (2021) *Securing new space: on satellite cyber-security* (Doctoral dissertation, University of Oxford), p. 26. Available at: https://ora.ox.ac.uk/objects/uuid:11e1b32a-8117-46b1-a0ce-9c485221d112 (Accessed at: 17 November 2023).

Pavur, J. and Martinovic, I., 2022. Building a launchpad for satellite cyber-security research: lessons from 60 years of spaceflight. *Journal of Cybersecurity*, *8*(1), p.tyac008, https://doi.org/10.1093/cybsec/tyac008.

Poirier, C. (2022) "The war in Ukraine from a space cybersecurity perspective", *European Space Policy Institute Report*, 84, pp. 5-10. Available at: https://www.espi.or.at/wp-content/uploads/2022/10/ESPI-Report-84.pdf (Accessed at: 15 November 2023).

Pultarova, T. (2023) "GPS satellites threatened more by mild solar storms than monster sun flares", *Space.com*, 10 August. Available at: https://www.space.com/mild-solar-storms-threat-gps-satellites (Accessed at: 17 November 2023).

Rafikasari, A., Sumarlan, S. and Swastanto, Y. (2020) "Challenges and opportunities in strengthening ASEAN space technology cooperation", *The Indonesian Journal of Southeast Asian Studies*, *3*(2), pp.173-187.

Raj, A. (2023), 'Why internet from satellites in space work best in Southeast Asia', *TechWire Asia*, 23 June. Available at: https://techwireasia.com/2023/06/why-internet-from-satellites-in-space-work-best-in-southeast-asia/ (Accessed at: 15 November 2023).

Rajagopalan, R. P. (2019) *Electronic and cyber warfare in outer space*. Geneva: UNIDR. Available at: https://unidir.org/wp-content/uploads/2023/05/electronic-and-cyber-warfare-in-outer-space-en-784.pdf (Accessed at: 16 November 2023).

Rome, N. (2021) "European Militaries Join the U.S. in Space", *Georgetown Security Studies Review*, 7 April. Available at: https://georgetownsecuritystudiesreview.org/2021/04/07/european-militaries-join-the-u-s-in-space/ (Accessed at: 15 November 2023).

Sarma, N. (2019). 'Southeast Asian space programmes: Capabilities, Challenges and Collaborations', *ORF Special Report*, 82, pp. 13-15. Available at: https://www.orfonline.org/wp-content/uploads/2019/03/ORF_SpecialReport_82_SEA-Space.pdf (Accessed at: 15 November 2023).

Satariano, A., Reinhard, S., Metz, C., Frenkel, S. and Khurana, M. (2023) "Elon Musk's Unmatched Power in the Stars", *The New York Times*, 28 July. Available at: https://www.nytimes.com/interactive/2023/07/28/business/starlink.html (Accessed at: 17 November 2023).

Schogol, J. (2023) "Army combat advisors testing military version of Elon Musk's Starlink", *Task & Purpose*, 9 October. Available at: https://taskandpurpose.com/news/army-sfab-starshield-spacex-elon-musk/ (Accessed at: 17 November 2023).

Scholl, M. and Suloway, T. (2021) "Introduction to cybersecurity for commercial satellite operations", *National Institute of Standards and Technology, Tech. Rep*. DOI: 10.6028/NIST.IR.8270

Space Foundation (2022) "Space Foundation and Singapore Space & Technology Limited Sign Partnership Agreement", 21 September. Available at: https://www.spacefoundation.org/2022/09/21/space-foundation-singapore-space-technology-limited-partnership-agreement/ (Accessed at: 17 November 2023).

Srivastava, M., Schwartz, F. and Sevastopulo, D. (2023) "China building cyber weapons to hijack enemy satellites, says US leak", *Financial Times*, 21 April. Available at: https://www.ft.com/content/881c941a-c46f-4a40-b8d8-9e5c8a6775ba (Accessed at: 17 November 2023).

Stokes, M.A., Alvarado, G., Weinstein, E. and Easton, I. (2020). "China's Space and Counterspace Capabilities and Activities", *Project 2049 Institute and The U.S.-China Economic and Security Review Commission*, 30 March, pp.21-93. Available at: https://www.uscc.gov/sites/default/files/2020-05/China_Space_and_Counterspace_Activities.pdf (Accessed at: 16 November 2023).

Strohmier, H., Stoker, G., Vanajakumari, M., Clark, U., Cummings, J. and Modaresnezhad, M. (2022) Cybersecurity Maturity Model Certification Initial Impact on the Defense Industrial Base. *Journal of Information Systems Applied Research*, *15*(2), pp. 22-25.

Symantec (2019) *Thrip: Ambitious Attacks Against High Level Targets Continue*. Available at: https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/thrip-apt-south-east-asia (Accessed at: 14 November 2023).

Takahashi, R. and Funabiki, N. (2022). Concrete Recommendations for Space Development in Non-spacefaring Countries. *ASEAN Space Programs: History and Way Forward*, pp.165-176.

Tan, H. (2023), 'Even Chinese companies are moving supply chains out China to avoid geopolitical risks. Here are the 6 places they're heading for instead.', *Business Insider*, 22 April. Available at: https://www.businessinsider.com/chinese-supply-chains-moving-companies-outside-mainland-2023-4 (Accessed at: 15 November 2023).

Thales Group (2023) *Thales seizes control of ESA demonstration satellite in first cybersecurity exercise of its kind*, 25 April. [Press release]. Available at: https://www.thalesgroup.com/en/worldwide/security/press_release/thales-seizes-control-esa-demonstration-satellite-first (Accessed at: 17 November 2023).

The Economist (2023) 'Keeping tabs on China's murky maritime manoeuvres', 15 August. Available at: https://www.economist.com/china/2023/08/15/keeping-tabs-on-chinas-murky-maritime-manoeuvres (Accessed at: 15 November 2023).

U.S. Department of Defense (2021) *Strategic Direction for Cybersecurity Maturity Model Certification (CMMC) Program,* 4 November. [Press release]. Available at: https://www.defense.gov/News/Releases/Release/Article/2833006/strategic-direction-for-cybersecurity-maturity-model-certification-cmmc-program/ (Accessed at: 17 November 2023).

U.S. Department of State (2023) *Joint Statement on U.S.-Singapore Space Dialogue,* 12 October. [Press release]. Available at: https://www.state.gov/joint-statement-on-u-s-singapore-space-dialogue/ (Accessed at: 17 November 2023).

UK Space Agency (2020) "Cyber Security Toolkit: Guidance on cyber security for space assets.", 19 May. Available at: https://assets.publishing.service.gov.uk/media/5ec298a3e90e071e2f955ebc/Space_cyber_toolkit_final_v4.pdf (Accessed at: 17 November 2023).

Umeda, K. (2023) "Satellite constellations taking on greater role in Japan's security", *The Japan Times*, 21 September. Available at: https://www.japantimes.co.jp/commentary/2023/09/21/japan/satellite-constellations-security/ (Accessed at: 17 November 2023).

UNOOSA (2023) "United Nations Office for Outer Space Affairs: Status of International Agreements Relating to Activities in Outer Space as at 1 January 2023". Available at: https://www.unoosa.org/res/oosadoc/data/documents/2023/aac_105c_22023crp/aac_105c_22023crp_3_0_html/AC105_C2_2023_CRP03E.pdf (Accessed at: 20 January 2024).

Upadhyay, S. N. (2023) 'Meet the Amazon for Space', *Analytics India Magazine*, 26 September. Available at: https://analyticsindiamag.com/meet-the-amazon-for-space/ (Accessed at: 15 November 2023).

Vasquez, C. (2023) "How a hacking crew overtook a satellite from inside a Las Vegas convention center and won $50,000" *CyberScoop*, 16 August. Available at: https://cyberscoop.com/mhackeroni-hackasat-space-def-con/ (Accessed at: 17 November 2023).

Verma, B. S. (2023) 'Rifle to Rafale: Space for defence', *ORF*, 22 September. Available at: http://20.244.136.131/expert-speak/rifle-to-rafale-space-for-defence (Accessed at: 15 November 2023).

Verspieren, Q. (2021) "Establishment of Space Operations Squadron at the Japan Air Self-Defence Force in 2020: current status and future prospects", *The Advanced Maui Optical and Space Surveillance Technologies* (*AMOS*) *Conference*, 22(3), pp. 1886-9. Available at: https://amostech.com/TechnicalPapers/2021/Poster/Verspieren.pdf (Accessed at: 17 November 2023).

Verspieren, Q. (2022) "Comparison of Established ASEAN Space Programs and Lessons Learned", *ASEAN Space Programs: History and Way Forward*, pp.125-127.

Verspieren, Q. (2023) 'ASEAN Space Programmes: Navigating Regional Rivalries', *RSIS Commentaries*, 008-23, pp. 1-2. Available at: https://www.rsis.edu.sg/rsis-publication/rsis/asean-space-programmes-navigating-regional-rivalries/ (Accessed at: 15 November 2023).

Weeden, B. and Samson, V. (2023) *Global counterspace capabilities: An Open Source Assessment*, Washington, DC: Secure World Foundation, pp. 13.06-13.12. Available at: https://swfound.org/media/207567/swf_global_counterspace_capabilities_2023_v2.pdf (Accessed at: 17 November 2023).

Wesler IV, W. (2016) 'The Democratization of Space', *RAND*, 28 March. Available at: https://www.rand.org/pubs/commentary/2016/03/the-democratization-of-space.html (Accessed at: 15 November 2023).

Wheeler, W.A., Cohen, N., Betser, J., Meyers, C., Snavely, W., Chaki, S., Riley, M. and Runyon, B. (2017) "Cyber resilient flight software for spacecraft", In *AIAA SPACE and Astronautics Forum and Exposition* (p. 5305). https://doi.org/10.2514/6.2018-5220.

Willbold, J., Schloegel, M., Vögele, M., Gerhardt, M., Holz, T. and Abbasi, A. (2023) May. "Space Odyssey: An Experimental Software Security Analysis of Satellites", *2023 IEEE Symposium on Security and Privacy (SP),* pp. 14-17. DOI: 10.1109/SP46215.2023.00131.

A Privacy-Compliant Process for Digital Forensics Readiness

Gabriel Shoderu, Stacey Baror and Hein Venter
University of Pretoria, South Africa

go.shoderu@gmail.com
stacey.baror@up.ac.za
hein.venter@up.ac.za

Abstract: This research paper examines the issue of privacy compliance in digital forensics readiness, specifically in relation to the breach of confidentiality during analysis of collected data on users. This is a problem because the collection and analysis of digital evidence during these investigations can have a potential impact on individuals' privacy rights. The study's methodology involves a literature review of relevant research, an analysis of privacy regulations, and a case study of a real-world digital forensics investigation. The main findings of the study indicate that organizations need to develop and implement robust privacy measures and data protection policies to ensure that their digital forensics readiness efforts are privacy compliant and do not compromise user privacy. Some examples of why this is necessary are provided in the research to address these privacy compliance issues, this study proposes a measure, this measure implements technical safeguards to protect user data and maintains its confidentiality. By implementing the proposed measures, organizations can maintain their digital forensics readiness while also protecting user privacy.

Keywords: Digital forensics readiness, Information privacy, Digital forensics investigation, Confidentiality, Digital forensics

1. Introduction

In today's digital age, cybercrime has become increasingly rampant and complex (Anderson *et al.*, 2019), posing significant threats to individuals, businesses, and organizations. To combat such criminal activities, digital investigative techniques are widely used to track and identify suspects. However, as the volume and severity of cybercrime continue to escalate, so do the challenges associated with conducting digital forensics investigations while maintaining privacy and data protection. One innovative solution to this problem is the development of digital forensic readiness which was first introduced by Tan (Tan, 2001), which takes a proactive approach to identifying and addressing digital evidence before any criminal activity occurs. This approach involves implementing policies, procedures, and technologies to facilitate the timely collection, preservation, and analysis of digital evidence in a way that is both effective and privacy compliant. Despite the many benefits of digital forensics readiness, there are still privacy concerns to be addressed (Yaacoub *et al.*, 2021; Anon, n.d.). These concerns centre on issues such as the collection and use of personal data, the retention and storage of digital evidence, and the potential for abuse of investigative powers. As such, any digital forensics readiness process must be designed and implemented with careful consideration of privacy and data protection regulations, ensuring that the rights and freedoms of individuals are respected and protected throughout the investigative process. Although digital forensics readiness ensures the quality and availability of evidence, it also involves collecting evidence in advance of a crime, which can be used to prosecute suspects anytime.

However, the problem is that the collection and analysis of digital evidence during these investigations can have a potential impact on individuals' privacy rights. For example, an organization investigates an employee for theft of confidential information, releases information to the media and employees, damaging the employee's reputation. The investigation clears the employee of all charges, but the damage is already done, and the information that was released about them cannot be taken back. The employee may face difficulty regaining the trust of their colleagues, clients, and competitors, and their career prospects may be negatively impacted. Therefore, this study aims to create a model that enhances the digital forensics readiness process in such a way that it is privacy compliant, while still maintaining the necessary level of investigative rigor.

The breach of confidentiality can lead to ethical concerns, as users have a right to privacy and expect their personal information to be always protected. The breach of confidentiality of user information can result in a damaged reputation of an organization, as users may lose the trust in their ability to protect their personal information. It may lead to non-compliance with privacy laws and could lead to financial penalties for organizations. In addition, collecting and analysing user information during digital forensics investigations can violate individuals' privacy rights, especially if sensitive or personal data is involved. This can be illegal under some acts and regulations in certain countries, such as the POPIA act. It is important that organizations have measures in place to protect user privacy, while still maintaining the necessary level of digital forensics readiness to detect, respond to, and investigate digital incidents.

This initial section, Section 1, serves as an introductory segment, setting the stage by defining the research problem and outlining the motivation behind the study. Subsequently, Section 2 delves into the background, providing essential context on digital forensic readiness and the regulatory framework surrounding user privacy. In Section 3, a model is developed, incorporating the most pertinent practices and techniques identified during the literature review. Section 4 introduces a comprehensive case scenario. The research further reviews related literature in Section 5, comparing the work of other authors with the proposed model. Finally, Section 6, the conclusion, provides a summative overview of the research and its contributions, offering insightful recommendations for future investigations in this field. The subsequent section delves deeper into the background literature.

2. Background Literature

This section delves deeper into the relevant literature on digital forensics to provide a more comprehensive understanding of the challenges and opportunities in this field and a background on user privacy regulations.

2.1 Background on Digital Forensics

The field of forensics initially focused on traditional forms of evidence collection and analysis such as fingerprints, DNA, and other physical evidence. As technology evolved and became more integrated into our daily lives, digital devices and systems started to play a more significant role in criminal investigations. This led to the emergence of computer forensics as a specialized field within forensics that dealt specifically with digital devices and systems such as personal computers. According to Daniel (Daniel and Daniel, 2011), digital forensics originated from computer forensics, which initially focused on personal computers. However, as the scope of computer forensics expanded to encompass various fields and practices, the term became limited. Thus, the term digital forensics was adopted to reflect the broader range of devices and systems that require forensic investigation.

Digital forensics is a branch of forensic science (Andre, 2018), that uses a scientifically derived and proven methods towards a distinct and definite process of deriving digital evidence from a digital source. The purpose of following these distinct processes is to facilitate or further the reconstruction of events as a result of criminal or non-criminal action. These processes are preservation, collection, validation, identification, analysis, interpretation, documentation, and presentation(Hein Venter, n.d.) This field encompasses multiple disciplines (Omeleze and Venter, n.d.; Hein Venter, n.d.), including computer science, cybersecurity, information security, conventional forensics, law, ethics and more. Computer science plays a crucial role in digital forensics, as both cybersecurity and forensics are built on computer-related concepts. While forensics is a subset of cybersecurity, it can also stand alone and be intertwined with cybersecurity. It also involves various fields including criminalistics, pathology, odontology, entomology and many more. (Rowlingson, 2004; Hein Venter, n.d.)

These various disciplines and fields must work together in order to uncover and analyse evidence through investigation. By utilizing a multidisciplinary approach, digital forensics can uncover evidence that may be hidden or difficult to find and use that evidence in legal proceedings. This uncovered evidence must be accurate, admissible, authentic, complete, reliable, believable and verifiable. (H. S. Venter, n.d.)

Digital forensic investigations typically focus on gathering evidence after the decision to initiate an investigation. However, the success of a digital forensics investigation relies on the availability and discovery of relevant evidence. If the necessary evidence exists and is located during the investigation, it can be used in charging and prosecuting a suspect. If the required evidence does not exist, it may impede the ability to press charges and carry out prosecution. This is where DFR comes in (HS Venter, n.d.), DFR attempts to gather the necessary evidence before an incident takes place in a bid to have more sound evidence when an investigation needs to take place, the formal definition is given by Rowlingson.

According to Rowlingson (Rowlingson, 2004), Digital forensic readiness is the ability of an organization to maximise its potential to use digital evidence whilst minimising the costs of an investigation. There are two main objectives of digital forensic readiness (HS Venter, n.d.), they are to:

- Maximize an environment's ability to collect credible digital evidence.
- Minimize the cost of forensics during an incident response.

The foundational DFR processes delineated by ISO IEC 27043 encompass five process classes, readiness process, initialization process, acquisitive process, investigative process and concurrent process, the fifth being executed concurrently. All these processes work together in order to fulfil the objectives stated above.

The following section delves deeper into what privacy is, what the regulation behind privacy is, techniques to address privacy.

2.2 Background on Information Privacy

2.2.1 Information privacy

Information privacy pertains to the right of individuals to exercise control over the collection, use, and dissemination of their personal information. Pavlou (Paul A. Pavlou, 2011) describes it as the concept of regulating the acquisition and utilization of one's personal information. Information privacy is heavily regulated by laws and regulations as discussed in the next section, one very specific regulation in the south African context is the Protection of Personal Information Act (POPIA) regulation.

POPIA is a significant regulatory framework that has a direct impact on digital forensic investigations involving personal information. POPIA aims to safeguard the privacy and protection of personal data in South Africa. (South African Government Gazette., n.d.)

Under POPIA, organizations are required to comply with certain principles and obligations when collecting, processing, and storing personal information. These principles include accountability, purpose specification, data minimization, accuracy, security safeguards, and data subject participation.

In the context of digital forensic investigations, the provisions of POPIA impose additional considerations and limitations on the collection, analysis, and storage of personal information. Investigators must adhere to the principles of lawfulness, fairness, and transparency when handling personal data. They should ensure that the collection of personal information is justified, limited to the purpose of the investigation, and conducted with the necessary consent or legal basis.

Furthermore, POPIA grants data subjects certain rights, such as the right to access their personal information, request its correction or deletion, and object to its processing. These rights may impact the procedures and practices followed during digital forensic investigations, requiring investigators to handle personal data in a manner that respects the rights of individuals. By incorporating the principles and provisions of POPIA into digital forensic investigations, organizations can ensure a privacy-compliant approach while preserving the integrity and admissibility of digital evidence.

In this research we will be making use of block-chain technology (Mohanta *et al.*, 2019) to enforce privacy, the following subsection sheds some lights on blockchain technology.

2.2.2 Blockchain Technology And Information Privacy

Blockchain technology offers an enticing solution for enhancing information privacy, particularly in digital forensic investigations. It serves as a decentralized, tamper-resistant ledger for securely recording data. (Mohanta *et al.*, 2019; Nakamoto, n.d.). Each data entry forms an immutable link in a chain, making it resistant to alterations—ideal for safeguarding sensitive information and Digital Forensic Readiness (DFR).

In the realm of information privacy, blockchain's decentralized structure encrypts and disperses personal data across multiple network nodes, reducing data breach risks. Additionally, blockchain enhances data transparency and accountability through smart contracts and consensus mechanisms, allowing individuals greater control over their data(Paul A Pavlou, 2011).

In summary, blockchain's security, transparency, and accountability have the potential to revolutionize information privacy within digital forensics, enabling organizations to align with privacy regulations while enhancing digital forensic practices. The subsequent sections detail our conceptual model and its components, providing a roadmap for privacy-compliant digital forensics readiness.

3. High Level View of BlockTrace Model

This section presents a holistic model aimed to seamlessly integrate the DFR processes with robust privacy standards, thereby creating a new phase of investigative readiness and adherence, we call this model, the BlockTrace model.

As this section progresses, the primary concern is attempting to harmonize digital forensics investigations with privacy standards. This model emerges as a direct response to the pressing need to not only enhance DFR but also ensure that every investigative action adheres to privacy principles.

The journey is started with a detailed breakdown of the proposed model and its respective components. Built upon the solid grounds of information privacy principles, the model aims to harmonize investigative efficiency with the protection of individuals' privacy rights.

As we delve into this section, we will exclusively explore these novel components and their profound impact on improving DFR processes while upholding the highest standards of privacy protection.

Commencing with the higher-level view of the BlockTrace model as shown in Figure 1, we embark on an exploration of each component we contribute, beginning with the overview of the Privacy initialization process (PIP), labelled B followed by the Privacy check process, labelled F.

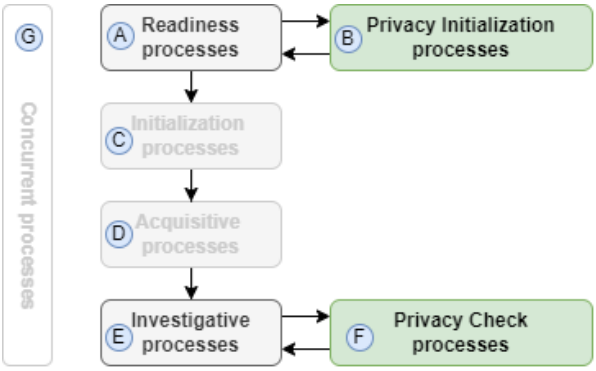


Figure 1: Adapting ISO/IEC 27043 (Labelled High-level view of BlockTrace Model)

3.1 Privacy Initialization Process (PIP)

The Privacy Initialization Process (PIP), a pioneering addition to the traditional DFR processes, represents a pivotal step towards integrating privacy-conscious practices into the very fabric of digital forensic readiness. This process establishes a solid foundation for safeguarding sensitive information while ensuring the legality and ethicality of digital investigations. At its core, the PIP focuses on fortifying data security and privacy before any investigative action occurs. A breakdown of the key steps for the PIP follows next:

- Compile User Data: This initial step involves gathering relevant user data that may be subject to investigation. This data serves as the starting point for subsequent privacy-enhancing measures.
- Encrypt User Data: The collected user data is encrypted using robust cryptographic techniques. Encryption provides an extra layer of security, rendering the data unreadable to unauthorized parties. This makes the data confidential.
- Hash Encrypted Data: The encrypted data undergoes hashing, creating a fixed-size representation or 'hash value.' This value serves as a digital fingerprint for the data, ensuring its integrity.
- Sign Hash with User Private Key: To establish authenticity, the hashed data is signed using the user's private key. This cryptographic signature acts as a seal, validating the origin of the data.
- Send Signed Hash to Ledger: The signed hash value is then stored in a blockchain ledger, leveraging the blockchain's inherent tamper-resistant properties. This step ensures an immutable record of the data's existence and its associated user.

These steps collectively constitute the PIP process. This process is seamlessly integrated into the sub-process group of the Readiness process class. More specifically, it will become an integral part of the implementation process group within the readiness process class, as shown in Figure 2. This systematic integration ensures that stringent privacy standards are upheld within the broader DFR framework.

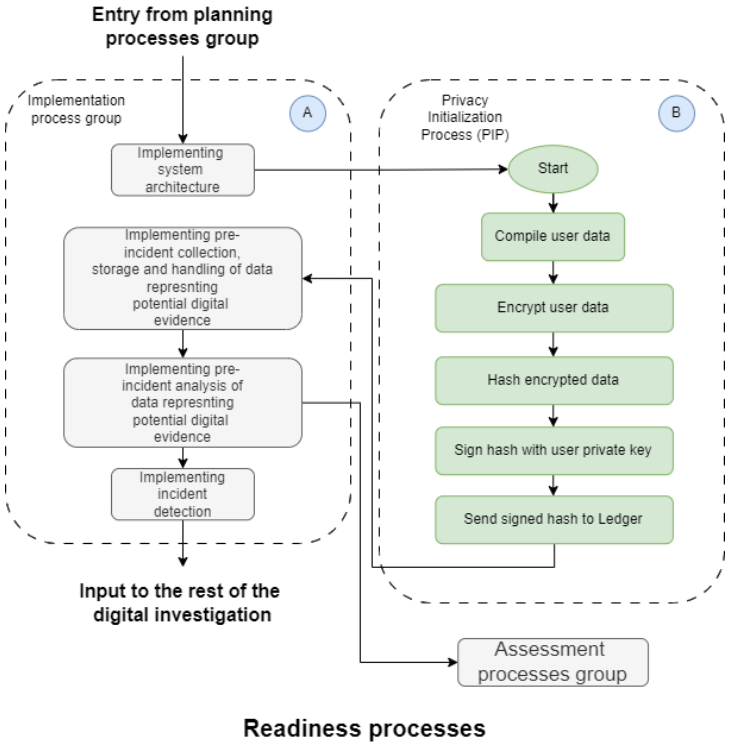


Figure 2: Integrated PIP into Readiness Process Class

3.2 Privacy Check Process (PCP)

Privacy Check Process (PCP) is an integral component of the BlockTrace model, this process plays a crucial role in ensuring that only authorized individuals have access to sensitive data, fostering a privacy-centric approach throughout the investigative journey.

The PCP is a distinctive addition to the investigative stage of the BlockTrace model. Its primary function is to validate whether an investigator possesses the necessary permission to access specific data for examination. This step ensures that data privacy and user rights are respected throughout the investigation process. Here's a breakdown of its key steps:

- **"Is the data accessible to the investigator?"**: At the outset of the process, there is an access authorization, the investigator's authorization level is verified. This involves confirming whether the investigator has the necessary permissions to access the data in question.
- **"Request access from data owner"** (expanded in next section): If the investigator's authorization is insufficient or unclear, they initiate a request for access to the data from the data owner. This request outlines the purpose and scope of the investigation. There is a multi-level access request process which is elaborated upon in the subsequent section.
- **"Retrieve data"**: Once authorization is granted, the investigator retrieves the encrypted data from storage. This encrypted data is safeguarded by privacy-enhancing technologies implemented earlier.
- **"Decrypt the data"**: The retrieved data is decrypted using the investigator's private key. This decryption step converts the encrypted data into its original, readable format.

These combined actions collectively constitute the entirety of the PCP process. This process, seamlessly interwoven into the Investigative process class, forms an integral element of the investigative journey, as depicted in Figure 3. The strategic incorporation of the PCP underscores the significance of user control over access to their distinct information.

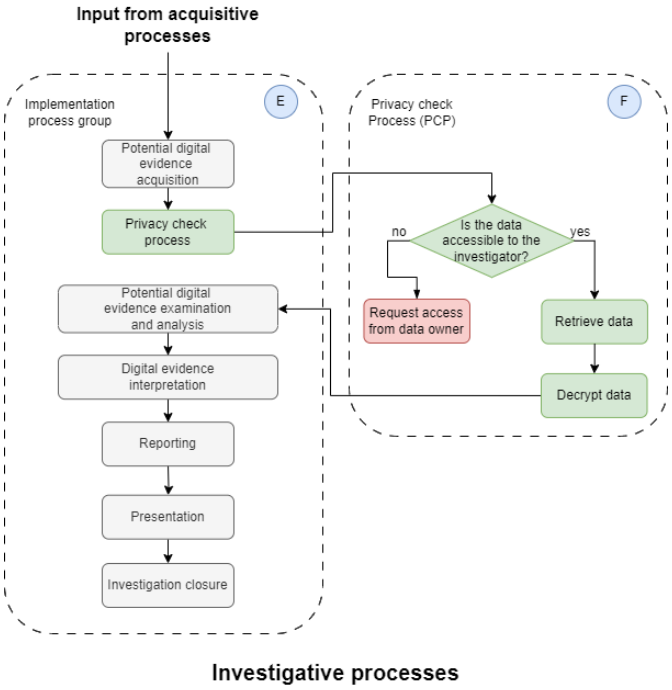


Figure 3: Integrated PCP into Investigative Process Class

3.3 Access Request Levels in Privacy Check Process (PCP)

In the Privacy Check Process (PCP), the access request phase plays a pivotal role in determining whether an investigator can proceed with data examination. This access request process is designed with multiple levels as shown in figure 4, ensuring a comprehensive approach to permissions and access control.

3.3.1 Level 1: Owner's access request

- The initial request for access to specific data is initiated by the investigator, directed to the data owner.
- The data owner holds the primary authority to grant or deny access based on their ownership rights and data stewardship responsibilities.

3.3.2 Level 2: Data custodian's evaluation

- If the data owner does not respond or provides an ambiguous response, the access request escalates to the data custodian.
- The data custodian is responsible for safeguarding and managing the data on behalf of the owner. They assess the investigator's request and may grant access if it aligns with the data's usage policies.

3.3.3 Level 3: Legal department review

- In cases where neither the data owner nor the custodian grants access, or when legal considerations are involved, the request proceeds to the organization's legal department.
- The legal department evaluates the request in accordance with legal requirements, compliance regulations, and the organization's data access policies. They may provide authorization or deny access based on these considerations.

3.3.4 Level 4: Escalation and mediation

- If access is still not granted after the legal department's review, the request enters the escalation and mediation phase.
- A designated mediation team or committee assesses the situation, seeking to reconcile any disputes or concerns among the involved parties. This phase aims to find a balanced resolution.
- Escalation and mediation serve as a final attempt to secure access while addressing privacy, legal, and organizational concerns.

- In cases where resolution remains elusive, the matter may be escalated to the jurisdictional legal system for potential civil litigation.

3.3.5 Access denied and investigation closure

- If, even after reaching the escalation and mediation phase, access remains denied, the event is locked, and the investigation is officially closed.
- This stringent multi-level access request process ensures that only authorized individuals can access sensitive data, while preserving data privacy and adhering to legal and organizational requirements.
- Closure in this manner is crucial to safeguarding sensitive information, as unauthorized access could potentially expose company secrets vital to perhaps maintaining a competitive edge in the market, leading to detrimental consequences, including business failure.

Implementing multi-level access requests via PCP bolsters data security and accountability in BlockTrace. This section introduces the BlockTrace Model, revolutionizing DFR processes. It harmonizes investigative efficiency with data privacy using advanced techniques and technology.



Figure 4: Access Request Levels in Privacy Check Process (PCP)

4. Case Scenario

In this case scenario, we illustrate the practical application of the BlockTrace model in a real-world digital investigation. Our protagonist is Alice, a digital investigator, who utilizes this model to probe a security breach involving Bob, an employee at a prominent tech organization, as depicted in Figure 5.

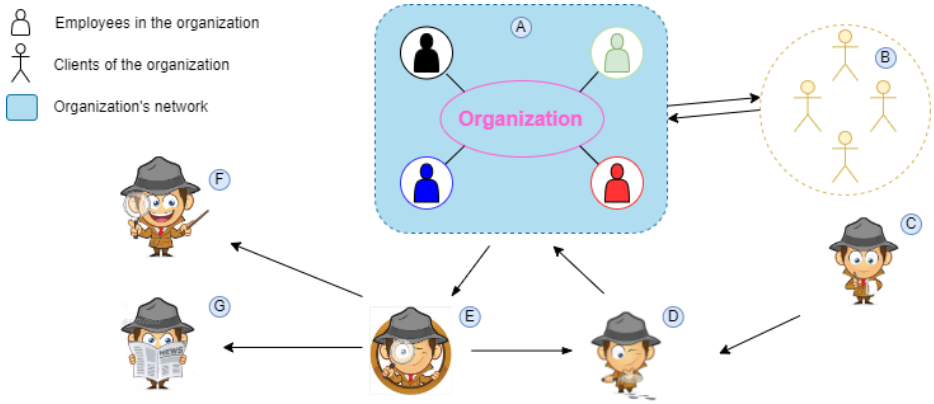


Figure 5: Case Scenario using Alice and Bob

As depicted in the illustration, starting from point A, the first essential step to enable a privacy-compliant investigation is the implementation of the BlockTrace model within the organization's network. The Privacy Initialization Process (PIP) ensures that all records of employees and their communications with clients, as shown in label B, are encrypted using their private keys.

Alice enters the scene at point C, where she's assigned to investigate a potentially fraudulent activity within the organization's network, indicated by point A. Her investigation leads her to Bob's computer, as portrayed in label D. To access particularly sensitive financial records, Alice initiates an access request through the Privacy Check Process (PCP), an integral component of the investigative process. In this request, she clearly outlines the investigation's purpose and scope.

This request initiates a multi-level access process, cycling through points A, D, and E. It begins with the Data Owner, Bob, who holds the initial decision-making authority. If Bob denies Alice's access request, it escalates to the Data Custodian, where it undergoes a comprehensive review to ensure alignment with the organization's established policies. If the request remains unapproved, it is subsequently escalated to the Legal Department for further scrutiny, ensuring alignment with legal and privacy requirements. Should access disputes persist, the next phase involves Escalation and Mediation, offering an opportunity to resolve access issues at a higher administrative level. As a final recourse for potential civil litigation, the Jurisdictional Legal System is prepared to intervene.

If access is granted at any point during this intricate process, the investigation can proceed, and all relevant data is decrypted. If not, the investigation is halted, and access is denied, marked by a locked event, as indicated in label G.

Upon concluding the investigation, Alice successfully uncovers substantial evidence of the alleged fraudulent activity, all the while preserving data privacy rights and adhering to strict legal standards, as depicted in label F. This scenario effectively underscores how the BlockTrace model safeguards data privacy and integrity in complex investigative endeavours, emphasizing the crucial integration of the Privacy Initialization Process (PIP) within the readiness process and the Privacy Check Process (PCP) within the investigative process.

5. Risks of Privacy Breach

This section highlights the concrete risks posed by privacy breaches in digital investigations. By offering specific examples, we aim to underscore the urgency of implementing robust safeguards for sensitive information.

5.1 Risks to Individuals

5.1.1 Identity theft

When personal information is inadequately protected during a forensic investigation, it becomes susceptible to theft. The compromised information becomes a gateway for malicious actors to assume false identities, leading to financial fraud and reputational damage.

5.1.2 Blackmail and extortion

Sensitive personal data exposed during an investigation can be used as leverage for blackmail or extortion. Individuals become vulnerable to coercion, with the potential release of sensitive information used as leverage for illicit gains.

5.1.3 Psychological impact

Individuals may suffer anxiety, stress, and loss of trust as their private lives are laid bare, impacting both personal and professional well-being. The knowledge that one's personal information has been exposed without consent can result in anxiety, loss of trust in digital systems, and a feeling of vulnerability.

5.2 Risks to Organizations

5.2.1 Legal and financial consequences

Privacy breaches exposes organizations to severe legal ramifications and financial penalties. Failure to protect privacy during forensic investigations may lead to hefty fines under regulations such as POPIA. The financial impact can affect organizations severely.

5.2.2 Reputational damage

A privacy breach can tarnish an organization's reputation, leading to loss of customer trust and confidence. This could lead to a long-lasting effect on customer retention and the ability to attract new business.

5.2.3 Intellectual property theft

In cases where a privacy breach exposes confidential business information, there is a risk of intellectual property theft. Competitors or malicious actors may gain access to trade secrets, proprietary processes, or innovative ideas, leading to a competitive disadvantage.

Privacy breaches pose intricate risks to organizations, spanning legal and financial implications, reputational damage, and the theft of intellectual property. Upholding privacy is not just a regulatory obligation but a strategic imperative, reinforcing resilience against these formidable threats.

6. Related Literature

In this section, we are bringing together insights from different types of studies from those delving into practical implementations to those probing the theoretical underpinnings. By understanding the valuable information in these pages, we equip ourselves with a comprehensive perspective, laying a foundation for the unique model we proposed. As we read through the pages of related literature, we not only acknowledge the collective wisdom of scholar preceding us but also strive to chart a path toward an integrated and privacy-conscious digital forensic future.

Frank Y.W. Law (Law *et al.*, 2011) and colleagues proposed a solution to privacy challenges in digital investigations. Their method suggests the application of encryption not only during data collection but also throughout data analysis. The core objective is to prevent digital investigators from accessing irrelevant data at any stage of the examination process.

Although this paper (Law *et al.*, 2011) focused on digital investigations in general, in the BlockTrace model, we have embraced and extended this encryption approach, integrating it to curb unauthorized access to specific data. This extension enhances data privacy and accountability in digital investigations.

Blockchain, originally conceived as the backbone of cryptocurrencies by Satoshi Nakamoto (Nakamoto, n.d.), has since matured into a multifaceted platform with far-reaching implications for data security and privacy. Its distinctive sequential and immutable structure makes blockchain exceptionally proficient at securing data and preserving its integrity.

Blockchain technology has found extensive applications across diverse domains (Mohanta *et al.*, 2019), driven by its capacity to ensure data confidentiality and integrity. Numerous research papers have explored the proposition and implementation of blockchain in various sectors. For instance, research has delved into the utilization of blockchain in the financial services industry (Treleaven *et al.*, 2017), while the IEEE has played a pivotal role in raising awareness about blockchain technology's potential in the healthcare sector (Mettler, 2016).

Within the context of the BlockTrace model, blockchain assumes a pivotal role as the linchpin of data privacy assurance. Through the seamless integration of blockchain technology, we not only bolster data security but also establish a robust underpinning for accountability within digital investigations. This amalgamation furnishes the means to curtail unauthorized access, meticulously trace data interactions, and unequivocally ensure the

immutability of digital evidence. Crucially, these advancements are achieved while upholding the fundamental privacy rights of all individuals involved.

This brief yet comprehensive review sets the stage for the proposed BlockTrace model's role in bridging the gaps identified in current literature, advancing the field of digital forensics while safeguarding individual privacy This brief yet comprehensive review sets the stage for the proposed BlockTrace model's role in bridging the gaps identified in current literature, advancing the field of digital forensics while safeguarding individual privacy.

7. Evaluation

The work presents several notable strengths. First and foremost, it effectively prioritizes data privacy by incorporating the Privacy Initialization Process (PIP) into the Digital Forensic Readiness (DFR) framework. This proactive approach ensures that data security is considered from the outset of an investigation. Furthermore, the integration of blockchain technology enhances data integrity within the BlockTrace model.

The blockchain's tamper-resistant nature provides a robust foundation for preserving the integrity of digital evidence. The introduction of the Privacy Check Process (PCP) adds a multi-level access control mechanism. This significantly contributes to safeguarding sensitive data and ensuring compliance with privacy regulations. The case scenario provides a practical example of how the BlockTrace model can be applied in complex digital investigations. This real-world applicability underscores the effectiveness of the model.

Lastly, the model's ability to seamlessly integrate privacy and security practices into existing DFR processes is a notable strength. It streamlines investigative procedures while respecting individual data privacy rights.

Several limitations should be acknowledged. The multi-level access control process introduced by the model can be intricate and time-consuming, potentially impacting investigation timelines.

Implementing blockchain and cryptographic techniques can be resource-intensive, necessitating investments in hardware and personnel training.

User training is essential, as users need to become proficient in navigating the model's new privacy and security features. This learning curve may pose initial challenges during adoption.

To enhance the model's effectiveness, several improvements could be considered. Streamlining the multi-level access control process could strike a better balance between security and efficiency, potentially optimizing the Privacy Check Process (PCP).

Exploring advanced privacy technologies such as homomorphic encryption and differential privacy might further enhance data privacy without compromising investigative efficiency.

The development of user-friendly interfaces for investigators could facilitate interaction with the BlockTrace model, improving its adoption and user-friendliness.

Expanding the range of case scenarios to cover a broader spectrum of investigative contexts will provide a more comprehensive understanding of the model's practicality.

8. Conclusion

In conclusion, the BlockTrace model, introduced in this paper, represents a significant step forward in enhancing the digital forensic readiness (DFR) processes while meticulously preserving data privacy. The model has successfully addressed the imperative need for a robust framework that harmonizes investigative efficiency with the protection of individual privacy rights. The Privacy Initialization Process (PIP) and Privacy Check Process (PCP) have been seamlessly integrated into the DFR framework, showcasing their pivotal roles in safeguarding data integrity and ensuring privacy compliance. The suggested future improvements in the evaluation aims to address these concerns and further refine the model's suitability for various investigative scenarios. One of these approaches may offer the means to access and analyse sensitive data without exposing confidential organizational secrets. By enabling a more nuanced, privacy-focused approach, we can balance the demands of thorough digital investigations with the protection of critical proprietary information, thereby contributing to a comprehensive and responsible digital forensic landscape. Overall, the work has the potential to set new standards in the field of digital forensics by bridging the gap between efficient investigations and data privacy.

References

Anderson, R. *et al.* (2019) 'Measuring the Changing Cost of Cybercrime'.

Andre, A. (2018) 'Digital Forensics'. p. 373.
Daniel, Larry E. and Daniel, Lars E. (2011) 'Digital Forensics for Legal Professionals: Understanding Digital Evidence from the Warrant to the Courtroom'. *Digital Forensics for Legal Professionals: Understanding Digital Evidence from the Warrant to the Courtroom*, pp. 1–330. DOI: 10.1016/C2010-0-67122-7.
Diaz, C. and Gürses, S. (2012) 'Understanding the Landscape of Privacy Technologies 1'. p. 1.
DIGITAL FORENSIC READINESS IN ORGANIZATIONS: ISSUES AND CHALLENGES - ProQuest. Available at: https://www.proquest.com/docview/2034195913?pq-origsite=gscholar&fromopenview=true (Accessed: 13 October 2023).
Law, F.Y.W. *et al.* (2011) 'Protecting Digital Data Privacy in Computer Forensic Examination'. *2011 6th IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering, SADFE 2011*. DOI: 10.1109/SADFE.2011.15.
Mettler, M. (2016) 'Blockchain Technology in Healthcare: The Revolution Starts Here'. *2016 IEEE 18th International Conference on E-Health Networking, Applications and Services, Healthcom 2016*. DOI: 10.1109/HEALTHCOM.2016.7749510.
Mohanta, B.K. *et al.* (2019) 'Blockchain Technology: A Survey on Applications and Security Privacy Challenges'. *Internet of Things*, 8, p. 100107. DOI: 10.1016/J.IOT.2019.100107.
Nakamoto, S. 'Bitcoin: A Peer-to-Peer Electronic Cash System'. Available at: www.bitcoin.org (Accessed: 15 October 2023).
Omeleze, S. and Venter, H.S. *Digital Forensic Application Requirements Specifications Process Word Count=8590*.
Pavlou, Paul A. (2011) 'State of the Information Privacy Literature: Where Are We Now and Where Should We Go?' *MIS Quarterly: Management Information Systems*, 35(4), pp. 977–988. DOI: 10.2307/41409969.
Pavlou, Paul A. (2011) *State of the Information Privacy Literature: Where Are We Now And Where Should We Go?*
Rowlingson, R. (2004) 'A Ten Step Process for Forensic Readiness'. *International Journal of Digital Evidence Winter*, 2(3). Available at: www.ijde.org (Accessed: 15 October 2023).
South African Government Gazette. *Protection of Personal Information. [Online]*. Available at: https://www.gov.za/sites/default/files/gcisdocument/201409/3706726-11act4of2013popi.pdf (Accessed: 15 October 2023).
Tan, J. (2001) 'Forensic Readiness'. Available at: http://project.honeynet.org (Accessed: 13 October 2023).
Treleaven, P., Brown, R.G. and Yang, D. (2017) 'Blockchain Technology in Finance'. *Computer*, 50(9), pp. 14–17. DOI: 10.1109/MC.2017.3571047.
Venter, HS. 'DFR'.
Venter, Hein. *Digital Forensic Investigations Introduction to Digital Forensics*.
Venter, H.S. *Digital Forensic Investigations The Traditional Digital Forensic Investigation Process-Part B*.
Yaacoub, J.-P.A. *et al.* (2021) 'DIGITAL FORENSICS VS. ANTI-DIGITAL FORENSICS: TECHNIQUES, LIMITATIONS AND RECOMMENDATIONS A PREPRINT'.

Impact of Cyber Security Operations on Hardware Requirements for Stable and Workable Industrial Environments

Jussi Simola, Arttu Takala, Riku Lehkonen, Tapio Frantti and Reijo Savola
University of Jyväskylä, Finland

jussi.hm.simola@jyu.fi
arttu.h.takala@jyu.fi
riku.p.lehkonen@jyu.fi
tapio.k.frantti@jyu.fi
reijo.m.savola@jyu.fi

Abstract: Securing electricity distribution is one of the most important principles of the EU cyber security strategy. For example, European cyber security regulations, such as NIS2 (Network and Information Security Directive), CER (Critical Entities Resilience Directive), and Cyber Resilience Act (CRA) together aim to create a foundation and guidelines for international standards in various industries and the operation of critical infrastructure. Securing critical infrastructure is a common goal for Western operators. The new European Union (EU) directives bring new requirements to critical infrastructure administrators, device manufacturers and operators. Previously, member states have had responsibility for compliance with the directives, but they have been given freedom in the method by which they approach the requirements. Currently, member states' solutions are not always uniform, which has led to increased difficulties in coordination on a multi-national level. This, in turn, may lead to difficulties in coordination when responding to cybersecurity threats and attacks on critical infrastructure. The new regulation focuses on unifying the reporting between member states, reporting requirements of severe critical infrastructure events, and creating cybersecurity risk management procedures. In this study, we will provide a novel solution on how critical infrastructure administrators, device manufacturers, and operators may respond and become compliant with the new EU directives. To reach compliance and to enable the responsibilities that are required by the directive, the critical infrastructure devices and environment must have the capability to enable the responsible parties to identify, protect, detect, respond, and report. This sequence of actions is cyclical in nature since the identification of threats and vulnerabilities requires reports, which in turn requires data and detection. Our study focuses on the hardware requirements this causes on the manufacturing specifications, such as data collection and detection capabilities. The research belongs to the CSG project, and the purpose is to develop a governance model to minimize Operational Technology related risks and create a new standardized operating environment for the seamless utilization of energy solutions and industrial environment. The results of the study will be used in the analysis of requirements definitions in the OT environment.

Keywords: Operational technology, Security operations, Cyber resilience, Governance model

1. Introduction

The first research of the CSG (Cyber security governance of operational technology in sector-connected smart energy networks) project handled a research approach for investigating cyber security at the operational and technical levels. It focused on where to concentrate on OT-SOC-related cyber security research and how we aim to deploy a testbed to develop a governance model in the CSG project. This is the second research paper regarding the CSG project, and it concentrates on the cybersecurity-related strategic requirements, which reflect the measure of the security operational centers and the operational industry level.

The research belongs to the CSG project, where the purpose is to develop a governance model to minimize Operational Technology related risks and create a new standardized operating environment for the seamless utilization of energy solutions and industrial environment. The results of the study will be used in the analysis of requirements and new design process definitions in the OT-SOC environment where the Industrial Control Systems (ICS) are crucial operative factors in an industrial environment. The key results create the basis for standardization work and create essential development needs for industrial business operations to develop cyber security.

The purpose of the research is to find out and compare international and national level requirements for the management of future industrial cyber security in industrial companies and compare how the initial use case findings and observations from the testbed platform help define new operational technology functionalities.

As the energy sector is the biggest contributing factor in achieving the EU's climate goal of reducing emissions by at least 55% by 2030, the Commission proposes to act on the Plan for the accelerated digitization of the sector, which is necessary to ensure that the transition to renewable energy sources, networked mobility, smart

buildings, and a more integrated energy system, at the heart of which are consumers. (European Council of the European Union, 2023)

Extensive energy disruptions in the US and EU over the past years have shown the need for resilient and cyber-secure energy.

What European Union-level legislation demands from national actors on a strategic level is noteworthy. We must compare what new threats and risks can arise if the legislation is implemented or is not implemented at the national level and followed at the company level. Based on this preliminary assessment, a deeper threat and vulnerability assessment can be made at different levels.

The main research question concerns the formulation of hardware requirements for cyber security operations such as information sharing. Cybersecurity requirements that we have in different kinds of Security Operations Centers set challenges to different kinds of digital-based operations. The past and present Command and control functions that we have for example, in SCADA systems differ. Therefore, it is crucial in the OT environment to analyze crucial needs of data and information content. How can we get proper and useful information from the industrial environment, and what information is needed? The test-bed environment or laboratory is built to answer questions such as those. The laboratory will help design the process of the cybersecurity governance model for the industrial environment.

2. Challenges in Standardizing the Elements of the OT Environment

To achieve the cyber security requirements of the European Union, member states must create or re-organize the authorities responsible for supervisory tasks. The large EU-level aim is that the European cyber defense strategy must be implemented within the EU member countries to achieve common situational awareness (European Commission, 2020a).

According to the European Commission (2022b), EU Member States must, by using a risk assessment, identify critical entities that provide essential services for the maintenance of functions vital to society, economic activities, public health, and safety or the environment, and identify cases in which an incident would have significant disruptive effects on these essential services or when it would affect the national systems that safeguard the rule of law. To ensure the proper functioning of the internal market, two main goals were defined related to the Cyber Resilience Act (CRA) (European Commission, 2022): Create the conditions for the development of secure products with digital elements by ensuring that hardware and software products are brought to market with fewer vulnerabilities and ensure that manufacturers take security seriously throughout the product's life cycle; and create conditions in which users can take cybersecurity into account when choosing and using products with digital elements.

The following goals were set at the initial regulation meeting: To ensure manufacturers improve the safety of products containing digital elements from the design and development stage and throughout the entire life cycle. To ensure a consistent cybersecurity framework to facilitate compliance by hardware and software manufacturers. To increase the transparency of the security features of products containing digital elements and enable companies and consumers to use products containing digital elements safely (European Commission, 2022). The regulation set new requirements for cybersecurity in the industry environment. How does that affect the three selected/strategic, operational, and technical levels?

According to the (EU) 2022/2555 directive, we don't have cross-administrative authority that may perform supervisory activities between the different sectors. Reporting mechanisms have been organized mainly by sectors (European Parliament, 2022). Political governance is seen as the highest level of governance. The political level focuses on establishing processes, roles, and responsibilities for implementing the National Cyber Security strategies in their related policies. Dedicated cybersecurity agencies or departments within ministries are leading the political level (ENISA, 2022a)

2.1 Strategical Cybersecurity Requirements to the Industrial Stakeholders

According to the European Parliament (2022), companies must be divided into industry specific categories. Entities would be classified based on their importance. There will be two main categories, the essential and important entities. The directive requires companies to implement cybersecurity management as part of the other management strategies. The high-level critical sectors were Health, Transportation, Banking and financial market, Digital infrastructure, Water supply (Waste water), Energy, ICT service management, Public Administration, Space, and Digital service providers.

2.2 Operational Requirements - Continuity Management

Operational requirements are connected to business continuity management. Companies must conform in a manner where operational continuity management is verified as required by cyber security regulations. (European Commission, 2022; PECB, 2023) So-called national ISAC´s (Information Sharing and Analysis Centre) are communities that share information among their sector-based groups by using, for example TLP (Traffic Light Protocol) -mechanism, but it does not have joint cybersecurity elements to inform about cyber threats and vulnerabilities. However, the viable mechanism is under development (ENISA, 2023b).

2.3 Technical Requirements

NIS2 directive will set the cybersecurity-related requirements for critical cyber functions. Therefore, every organization must organize its core functions in a cyber-resilient way. Especially Cyber Resilience Act (CRA) sets new standards for industrial organizations at the technical level (European Commission, 2022).

3. Central Concepts

CSG (Cybersecurity Governance of operational technologies in the sector connected smart energy)

The main aim of the Business Finland-funded CSG consortium is to create a replicable governance model for the energy and industrial sector. Co-operative development work with business stakeholders and research organizations makes it possible to understand different industries better. The built test laboratory environment supports the main aim of the project.

The project consists of a cybersecurity governance model covering the industry operations, selected digitalized devices, used equipment, training employees, managing procedures, and information sharing process for operational technology that belongs on it. There will be three main governance levels: Strategic, Operational, and Tactical.

SOC & C2

In US basic military terminology, Command and Control (C2) is also described as the exercise of authority and direction by a properly designated commander over assigned and attached forces in the accomplishment of the mission (Department of Defense, 2021). Command and control systems are the facilities, equipment, communications, procedures, and personnel essential for a commander to plan, direct, and control operations of forces pursuant to the missions assigned (Department of Defense, 2021).

The origin of this practice is in military operations, but it has evolved and spread across multiple industries, which has led it to be sometimes called “universal C2”, due to the similarity of sub-functions. A well-known implementation of this is NASA mission control, due to the complexity of the missions (Maykish P.,2014) As network infrastructure developed, the apparent need for C2 in complex networked environments became apparent, and network-centric warfare’s prominence led to military operations employing C2 in the form of Network Operation Centers (NOC) (Kaliyaperumal L., 2021). As networked environments evolved, so did the need for its protection, which led to NOC monitoring evolving into Security Operation Centers (SOC), with the capability of intrusion detection and response through reactive monitoring. Currently, sophisticated SOCs can proactively monitor with automation Kaliyaperumal L., (2021), but implementing SOC at this level into OT environments is challenging due to environments’ differences.

Cyber resilience

Cyber resilience is a fundamental feature in the OT environment when the main pursuit is to develop a safe and stable supply chain. Business processes belong to business continuity management (PECB, 2023). Operational industrial environments need standardization because workable, seamless industrial processes make it possible to enhance business processes. Standardized process management by Business Continuity Management (BCM, 2023) and also cybersecurity-guided development of devices are crucial factors when the purpose is to maintain fluent supply chain operations.

The white house (2013) defines resilience as the ability to prepare for and adapt to changing conditions and withstand and recover rapidly from disruptions. Resilience includes withstanding and recovering from deliberate attacks, accidents, or naturally occurring threats or incidents. Achieving resilience requires accurate information and analysis of risks. “Resilient infrastructure assets, systems, and networks must also be robust, agile, and adaptable. Mitigation, response, and recovery activities contribute to strengthening critical infrastructure resilience” (The White House, 2013).

Governance

Governance describes a complex system, defining roles, responsibilities, processes, and relationships. Governance includes stakeholders from and covers the private sector, the public administration as well as civil society and spans different topics such as economic, social, and political priorities (ENISA, 2023a). Cyber Governance is the operation of decision-making processes that enhance and ensure "participation, transparency, and accountability in taking measures related to cyberspace together with the mechanism of international agreements, strategies, laws, measures, regulations, and standards that interlock in the best way" (Efe, A. & Bensghir, K. T., 2019).

Threat Information Sharing

Cyber threat information helps organizations identify, assess, monitor, and respond to cyber threats. Cyber threat information may consist of indicators of compromise, tactics, techniques, and procedures used by threat actors, suggested actions to detect, contain, or prevent attacks; and the findings from the analyses of incidents (NIST, 2020).

Indicators are technical artifacts or observables that suggest an attack is imminent or is currently underway or that a compromise may have already occurred. Indicators can be used to detect and defend against potential threats. Examples of indicators may include the Internet Protocol (IP) address of a suspected command and control server, a suspicious Domain Name System (DNS) domain name, a Uniform Resource Locator (URL) that references malicious content, a file hash for a malicious executable, or the subject line text of a malicious email message (NIST, 2016).

Threat Reports describe tactics, techniques, and procedures, actors, types of systems, and information such as threat-related information that provides more accurate decision support and situational awareness to an organization. Threat intelligence consists of threat information that has been aggregated, transformed, analyzed, interpreted, or enriched to provide an important context for decision-making processes (NIST, 2016).

Tactics, techniques, and procedures (TTPs) describe an actor's behavior. Tactics are high-level descriptions of behavior, techniques are detailed descriptions of behavior in the context of a tactic, and procedures are even lower-level, highly detailed descriptions in the context of a technique. TTPs could describe an actor´s tendency to use a specific malware variant, order of operations, attack tool, delivery mechanism (e.g., phishing or watering hole attack), or exploit. (NIST, 2016).

Regulation

As we have seen in Europe, stable development of the security environment is not the same anymore. The eastern neighbor of Finland has started a war that affects all over the world. Therefore, we must take a couple of steps in advance to achieve a more secure state or situation where we want to live in the future. It is not possible to achieve a fully secured future, but we can do things and measures that allow us to develop better overall situational awareness including cyber situational awareness. Therefore, European-level security agencies and public safety actors have created cyber security strategies that must be implemented with European-level cyber security strategies in order to achieve joint overall security. Most European Union member states also have full membership in NATO, which secures our living in a changing security environment.

Digitalized transformation of traditional infrastructures in the world requires preventive actions against cyber threat events. Nowadays, hybrid threats are more common than ever. Cyber-physical threats set challenges because we must detect weak threats and signals much earlier than in basic or simple threat atmospheres. The new European-level legislation set new requirements for the European Union member countries, its enterprises, companies, and individuals, including management level and in-practice level workers (European Commission, 2022). For example, the manufacturing process and used components of the digitalized equipment and devices will be more standardized. NIS2 Networking and Information Systems directive sets requirements for the companies and their strategic, operational, and technical functions. On the other hand, the Cyber Resilience Act (CRA) supports the goals of the NIS2, and NIS2 supports the aims of the CER Cyber Resilience directive. CRA sets requirements for the manufacturing process of digitalized products, industrial companies, and cyber security training methods for the personnel and management of security operations (European Commission, 2023c). The aim is to create common industry-specific cybersecurity requirements for the global business environment. The detected cyber incidents must be reported to the national cybersecurity agency and the national cybersecurity agency has to resend those vulnerabilities information to the European level security agencies. It is also crucial that the information will be shared to other EU member states and sector-specific actors.

3.1 Laboratory Environment

The developed testbed implements high-end OT hardware and software solutions commonly used in critical infrastructure that are connected to an IT network, which enables it to be operated, maintained, and surveilled remotely. Developed testbed enables cybersecurity-related use cases that target each component, their connecting interfaces, and specifically the addition of IT solutions in OT environments. Current IT related cybersecurity research highlights the importance of both host and network detection solution, which is tested in practice with use cases. For this, a SOC will be implemented, with sensors at different levels of intrusion, to demonstrate the required depth of access the detection sensors may require, to provide the protection against malicious activity required by the regulations. For example, a sensor at the network layer might notify SOC that the OT device is attempting to exfiltrate information outside of the allowed operating region, but it would not provide any information about the malicious internal activity of the OT device. In addition to this, the SOC will be further developed with machine learning solutions to detect more sophisticated attack patterns, which enables more sophisticated use cases. The developed testbed will in this way provide data to support the development of maturity levels and highlight the possible risks that different levels of maturity have when it comes to the implementation of IT in an OT environment.

4. Research Methods

The built testbed in a laboratory environment allows to research scenarios in a way that the created threat scenarios are possible to execute in practice. The devices, equipment, and software that we have gotten from stakeholders enable us to do tasks as we have planned.

Initial laboratory tasks will produce new data for the governance model and help to develop features and functionalities of the security operations center. It will bring added value to the IT-SOC features because there are missing signal-processing elements that enhance the formation of situational awareness functionalities. It is crucial to gather data such as weak signals from the industrial environment where the old-fashioned machines are. The critical is to identify the emergency point, where SOC and detection features must react. In Industrial Control Systems such as Scada, intrusion prevention or detection tools are relevant machines that we must have under the secured control of operations. We will apply a widely used design science research methodology used traditionally in software and system development. As Hevner (2004) has explained in Figure 1, design science research methodology is based on three iteration cycles that support each other's. On the left side, there is an environment where problem formulation has been done. We must understand the environment where we are doing our study (people, organizations, technology).

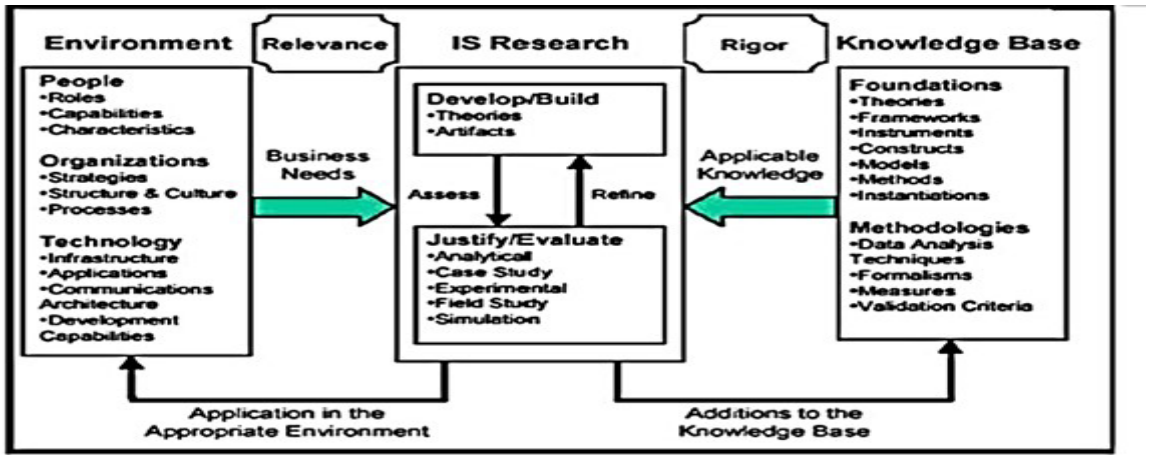


Figure 1: Information Systems Research Framework (Hevner et al. 2004)

Operations of the Design science research form by requirements from test environment that we have created. Case study research basement by Yin (2004) and the present knowledge base creates a core framework for the artifact and creates an added knowledge base. This output of the iterative process must be different than present system. The added value of the designed system must be in line with the original aim of the research.

As Yin (2004) states case studies solve real-life problems from real life. Exactly formed research questions such as "How does a hawk catch a mouse?" form the basis of the case study strategy. The research problem must be clearly defined. Why and when words are commonly used in research questions.

Nunamaker & al. (1991) argued that the systems development framework consists of five main stages: conceptual design, constructing the architecture of the system, analyzing the design, prototyping (may include product development), and evaluation.

We have also used The Delphi method which is suitable to make a relevant analysis of the testbed environment. "The Delphi method is an iterative process to increase consensus-building and, in the end, to have consensus among experts from an examined case." (Garson, 2012).

5. Results

Firstly, the systematic analysis of the study indicates the following things: There must be common overall situational awareness based on correct information as well as in the smallest industrial units and among the upper-level decision-makers. Horizontal and vertical situational awareness must be based on a common mental model and understanding of the current state of the overall security at the strategic, managing, operational, and tactical levels. That understanding or awareness should come for example, from the operational environment where industrial devices and components have crucial roles according to the continuity management and supply chain (NIST, 2022; United States Government Accountability Office, 2015).

If communication is not possible or if the information does not flow between the decision-maker level and the operational environment, potential obstacles prevent tackling cybersecurity threats. Fundamental problems in security management implementation processes set new challenges to implement and apply new cybersecurity technologies and tools (Simola, J., & Lehto, M. 2020).

Implementation processes at the technology/ technical level in practice set new challenges, and it must follow strategic cybersecurity goals at the national level and companies' business strategies. To achieve cyber resilience, it is important to find crucial factors that affect the whole supply chain in the plant's production line.

The testbed environment will show the weakest points that may generate vulnerable setpoints in the supply chain and information sharing. The testbed environment has been built for development work by using the idea of the DSR methodology, and we have opportunities to demonstrate real-life problems that occur in ICS environments.

The knowledge base Matrix developed by MITRE (2023) proves that ICS-related vulnerabilities are linked to the IT -environment in many ways. We are going to apply Use cases because one of the main aims is to design combined IT/OT Soc that make it possible to use proactive features in threat detection functions.

The Soc maturity is crucial in efficient proactive and prevention mechanisms. The threat scenario analysis illustrates that it is crucial to harmonize different kinds of soc features within the sector-based information-sharing groups. If there are different soc maturities in one sector-based (E.g., energy sector) sharing group, the weakest link governs the procedures. It is crucial and critical to set maturity assessment in a way that has been meant in cybersecurity scenarios state. If there is a cap on the activities of national authorities, it will create more vulnerabilities in industrial basements where hardware and software are connected to each other, and cyber threats are not detected in the way they had been.

Setting soc maturity at a common standard level also helps other countries and organizations to implement the operational systems and procedures to the required situation. Getting the required needs, in reality, is possible if sensor technology has been applied to all required cybersecurity stages.

5.1 Challenges Concerning Information Sharing in SOCs

In Europe, unlike in the US, Security Operation Centers are traditionally sold as a service from the private sector. To ensure the continuity of critical cybersecurity services, sharing sensitive information on higher levels of critical infrastructure and government level would benefit from SOCs not being established through competitive tendering as Figure 2 illustrates. Additionally, the government-affiliated SOCs would lower the threshold for private companies in critical infrastructure to share their vulnerable data, which helps develop more mature situational awareness (situational picture) as the data is collected to (higher levels) for further analysis. There should be horizontal and vertical possibilities to exchange data or information that is processed understandably.

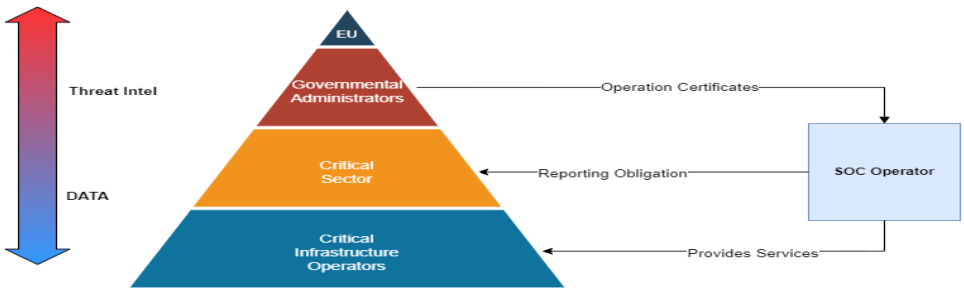


Figure 2: Information sharing pyramid

Since SOC as a service in the private sector is an established model, it can be used on the lower level of the framework as long as it is organized based on maturity and capability. This enables the SOC to report their findings informally and with the required depth and route based on the NIS2 directive.

5.2 Logic of Information Sharing Activities

In this sense, a SOC operator could provide their services to multiple critical infrastructure operators but would have to be affiliated to their operation sector for reporting purposes. To ensure continuity, the sector-based SOCs are specifically designed and built for that purpose rather than through competitive tendering. This additionally helps with competitive integrity, where a private SOC operator could provide services for sector SOC operations and operators within that sector, as Figure 2 illustrates. The collected data is analyzed, and overall countermeasures are provided to the sector-specific database, which helps sector operators through preventive countermeasures based on previous reports. Sector-specific reports are further reported to higher administrative level, which helps create a mature overall picture of situations between sectors and develop a unified countermeasure database, where other critical infrastructure sectors may search for possible preventive actions based on previous reports.

To draw a similarity to a well-known case with a malicious attack, while host device logs and network logs may have provided cohesive and matching data, additional sensors that are completely separated from its functional operating system could have showcased an anomaly. There is an advantage in the OT environment compared to the IT environment to use different kinds of sensor technology as Figure 3 demonstrates.

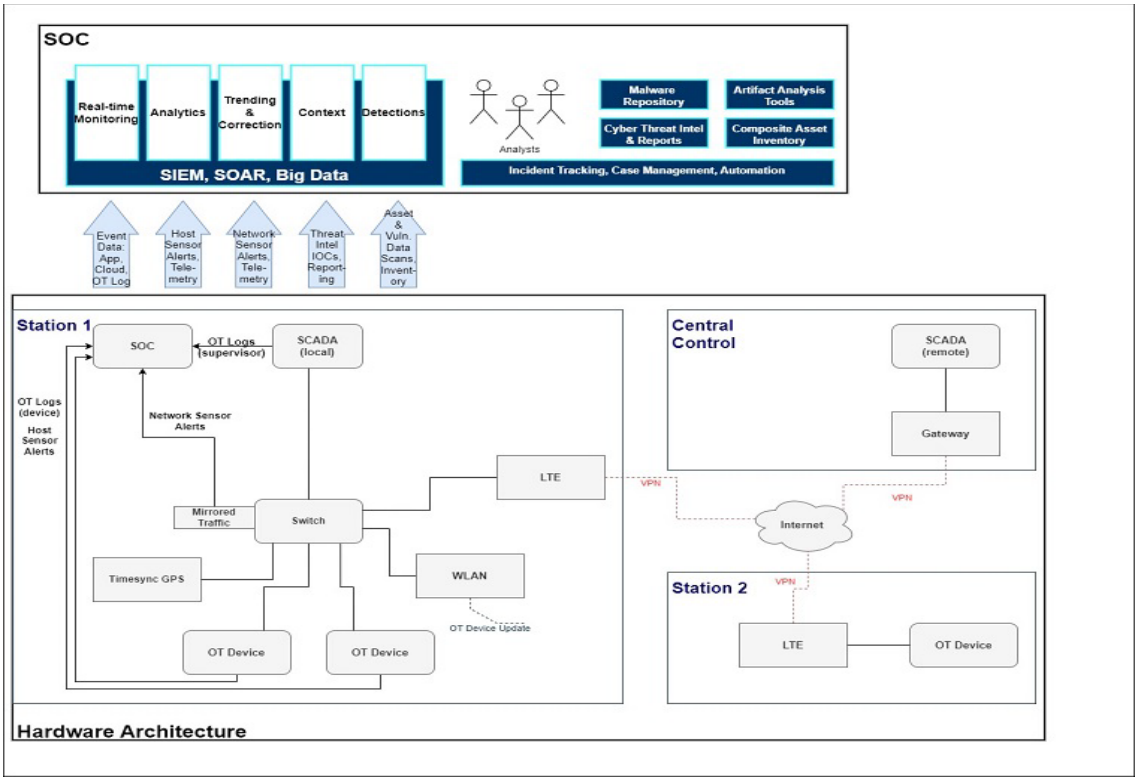


Figure 3: Hardware architecture and Soc environment (Modified from Knerler, K & al.2022)

Figure 3 demonstrates how developed testbed collects event data in the form of OT logs from OT devices and SCADA software, host sensor data from OT devices, and network sensor alerts based on the mirrored network traffic occurring through the switch. Alerts occurring from the data are audited for reporting and threat intel IOC purposes, which functions as an interface for the overarching reporting in the NIS2 directive. OT environment is unique compared to the IT environment, where asset and vulnerability data scans are seen as too disruptive, which is why only inventory documentation is seen as a possible solution. Deep scans are too harmful to the system. Scanning open ports doesn´t load that much. One important thing to measure will be how heavy scanning can be done. Twinned systems might be a solution for scanning, but they would require expensive hardware duplication. Twinned system would additionally help with critical system updates.

The multi-sensor system allows internal network traffic to be analyzed throughout its travel, allowing the SOC to alert of possible tampering midway through the operation. This leads to malware needing multiple devices, if not the whole system, under control to remain hidden while tampering.

Traffic mirroring is the initial step in network sensors. However, SOC is, in this case, connected to the internal network, and the attacker that controls the switch may notice the monitoring. Additionally, it only monitors the internal network from the switch's point of view, whereas advanced monitoring would be through data diodes, so the monitoring remains undetected, and possible signal or packet tampering between device connections may also be monitored.

The problem is that SOC operators can also be critical infrastructure operators. Threat reports must be in a concise and comprehensive format so that the information is not lost on the way. For example, if the IP is different for the attacker in different sectors, but the attack chain remains the same. The critical sectors must have their own cyber security capability to handle comprehensive and concise threat reports. National cybersecurity authority should transfer relevant information to different sectors. The energy sector does not know what devices other sectors have, so it is not possible to know if any IT/OT information is relevant to other sectors. Figure 4 demonstrates the problem formulation of different kinds of maturity levels in SOC´s.

Important things to note: On what scale are horizontal and vertical operations in terms of reporting? Where is any level of cyber security capability located? In other words, overlapping activities must be reduced and critical information threatening vital functions must be shared with the authorities regardless of the level of maturity.

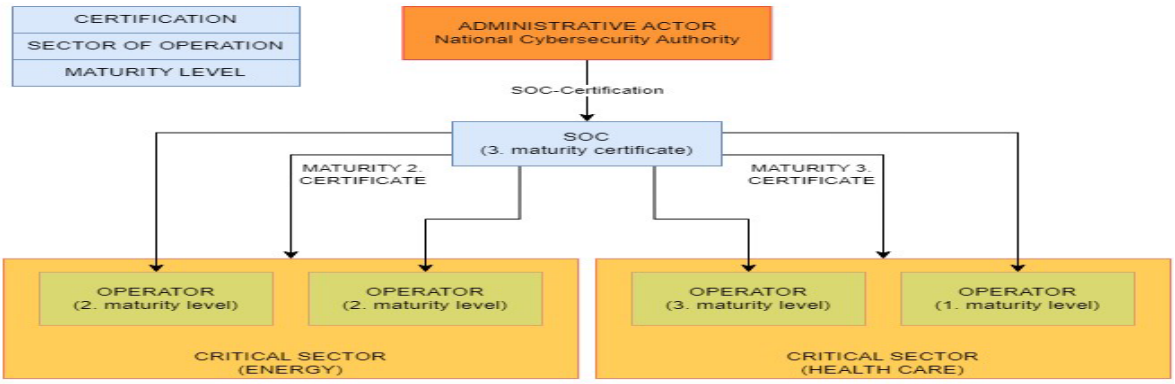


Figure 4: Maturity of the SOC´s

If the SOC activity is locked behind certificates on a sector-by-sector basis, a reporting system can be created. At the same time, a classification of SOC abilities is created, and operators are divided into different ability levels. For example, SOC has the 4th maturity for operations, but its know-how in the energy sector is only the 3rd maturity. In this case, operators in the energy sector can only buy protection of 1-3 maturities from SOC.

Cooperation between operators can be managed through their maturity levels. This enhances situational awareness and governance within a sector since a lower-maturity cybersecurity solution could undermine the network as a whole.

Additionally, there are differences between the countries; protection level may set soc operators in unequal situations from each other. Secure Continuity management sets its own requirements for a situation when operations are sold abroad.

For example, the SOC always reports the operator's problems to the operator's sector (e.g., X IP address is bad), from which the sector's other SOC operators or operators can obtain information for their own measures as

Figure 5 illustrates. Correspondingly, higher-level situational information must be kept, which is why the overall events of the sector are reported to the administrative operator.

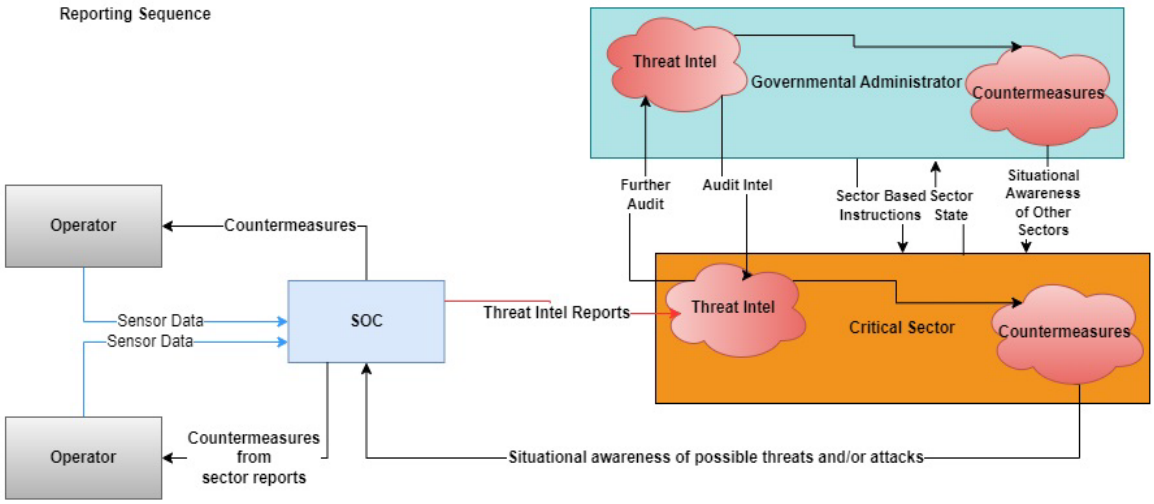


Figure 5: Reporting Sequence

The administrative actor also receives more extensive reports of cyber security events from SOCs to improve situational awareness.

6. Discussion and Conclusion

Ongoing implementation work of regulation is not a separate section from the development work of a joint strategic view of the Western cyber-attack prevention tendency. As we have seen, strategic aim and operational, tactical technology-based approach are not separate from each other. Designing a new governance model for the governance OT environment requires a coherent understanding of every awareness state. Political, strategic, operational, and technical/tactical levels must support each other to achieve common cyber situational awareness. Strategical level requirements face device-level requirements to notice deficiencies in equipment and justification for these. On the other hand, practical tests in a testbed environment produce requirements for future maturity aspects.

Standardized SOC features are crucial for the formation of joint situational awareness. Information-sharing mechanisms are crucial factors that affect supply chain management. Control processes must be formed in order to achieve real-time data from the factories. The information should be shared with other stakeholders by following standards and guidelines that are consistent with best practices. We must have at least two levels of information that industrial organizations of critical infrastructure should share within the sector-based groups. Overall cybersecurity maturity level is achievable only by implementing all required elements into the practice. That means the flow of information should achieve all required stakeholders at the time, information should be reported in an understandable way (consisting of relevant information) and it has to affect procedures and processes that are effective and efficient, preferably proactive. Shared information is not enough to achieve resilience in the cyber-physical ecosystem.

References

Department of Defense (2021). DOD Dictionary of Military and Associated Terms. U.S.

Efe, A. & Bensghir, K. T., (2022). cited in Savas S. & Karatas, S. Cyber governance studies in ensuring cybersecurity: an overview of cybersecurity, International Cybersecurity Law Review.

European Commission, (2020a). The EU's Cybersecurity Strategy for the Digital Decade. Brussels. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020JC0018

European Commission, (2020b). Proposal for a directive of the European parliament and of the Council on the resilience of critical entities. COM (2020) 829 final. 2020/0365 (COD). Brussels, 16 December.

European Commission, (2022c). Proposal for a regulation on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020.

European Parliament, (2022). Directive 2022/2555 Network and information security (NIS2).

European Council of the European Union, (2023). European Green Deal https://www.consilium.europa.eu/en/policies/green-deal/

ENISA, (2023a). Building Effective Governance Frameworks for the Implementation of National Cybersecurity Strategies.

ENISA, (2023b). Information Sharing and Analysis Centers, ISAC in a box. https://www.enisa.europa.eu/topics/national-cyber-security-strategies/information-sharing/isacs-toolkit/view#build/governance-structure

Garson, G. D. (2012). The Delphi method in quantitative research. Asheboro, NC: Statistical Associates Publishers. Available from: https://faculty.chass.ncsu.edu/garson/PA765/delphi.htm, retrieved 23.10.2023.

Hevner A., March, S.T., Park, J., and Ram, S. (2004) Design Science in Information Systems Research. MIS Quarterly.

Kaliyaperumal L.N. (2021). The Evolution of Security Operations and Strategies for Building an Effective SOC. https://www.isaca.org/resources/isaca-journal/issues/2021/volume-5/the-evolution-of-security-operations-and-strategies-for-building-an-effective-soc

Knerler, K., Parker I., Zimmerman C. (2022). 11 Strategies of a World-Class Cybersecurity Operations Center. MITRE Corporation.

Maykish. P., (2014). C2 Rising: A Historical View of Our Critical Advantage in Air & Space Power Journal.

MITRE (2023). "ATT&CK Matrix for Enterprise," [online], https://attack.mitre.org/.

NIST (2016). Special Publication 800-150 Guide to Cyber Threat Information Sharing.

NIST (2022). SP 800-161 Rev. 1 Cybersecurity Supply Chain Risk Management Practices for Systems and Organizations.

PECB, (2023). ISO/IEC 27001 vs, ISO 22301 vs. ISO 31000: What you need to know? https://pecb.com/past-webinars/isoiec-27001-vs-iso-22301-vs-iso-31000-what-you-need-to-know

Nunamaker, J., Minder Chen J. R. and Purdin T., "Systems Development in Information Systems Research," (3), pp. 89-106, 1991.

Simola, J., & Lehto, M. (2020). National cyber threat prevention mechanism as a part of the E-EWS.In B. K. Payne, & H. Wu (Eds.), ICCWS 2020 : Proceedings of the 15th International Conference on Cyber Warfare and Security (pp. 539–548). Academic Conferences International. https://doi.org/10.34190/ICCWS.20.106

United States Government Accountability Office, (2015) Critical Infrastructure Protection. Sector-Specific Agencies Need to Better Measure Cybersecurity Progress.

White House, (2013). Presidential Policy Directive/PPD-21 -- Critical Infrastructure Security and Resilience

Yin, R.K. (1994). Case Study Research: Design and Methods, 2nd edn. Sage Publishing, Thousand Oaks.

Hardware Sequence Combinators

Stephen Taylor[1] and Gunnar Pope[2]
[1]Thayer School of Engineering, Dartmouth College, Hanover NH, USA
[2]BitStory.AI, LLC, USA

stephen.taylor@dartmouth.edu

Abstract: Recent advances in formal methods for constructing parsers have employed the notion of *combinators*: primitive elemental parsers with well-defined methods for combining them in sequences or through choice. This paper explores the subtleties associated with leveraging *sequence combinators* to produce compact, custom hardware traffic validators. This involves a fully automated process that takes as input a formal grammar specifying message formats and produces a parsing circuit capable of validating traffic headers and payload content. The resulting circuit is deployed through network *guard* appliances that employ Field Programmable Gate Array (FPGA) devices, or alternatively, within the on-chip FPGA associated with System-on-Chip (SoC) devices, such as the Xilinx UltraScale MPSoC. Each guard appliance acts as a hidden "bump-in-the-wire" that either forwards or drops individual packets based on the message parsing outcome, thereby hardening network segments against zero-day attacks and persistent implants. Guards may operate on a wide variety traffic protocols and formats including TCP/IP, CAN/J1939, or MIL-STD-1553. The central step in parser construction is to build a collection of standard shift/reduce parsing tables that can be employed by a push-down automata to check each byte in a message. Typically, these tables are sparse, resulting in excessive use of FPGA circuit resources to represent them. By leveraging sequence combinators, along with other optimizations, we have been able to produce highly compact representations that can reduce table size by up to 95% for non-trivial grammars. Depending on the grammar, this translates directly into FPGA resource reductions. The reductions now make it viable to implement complex parsers on small, inexpensive FPGA's, or alternatively combine parsers with encryption and encapsulation to enhance guard capabilities.

Keywords: Parsing, LALR grammar, Traffic validation, FPGA, Bison, Hammer

1. Background

Air, space, and ground vehicles, like many other embedded systems, increasingly rely upon standard networking technologies to link embedded control systems with sensors, actuators, and human machine interfaces through industry standard buses (e.g., CAN/J1939, MIL-STD-1553, USB) and communications interfaces (e.g., Gigabit Ethernet, OpenVPX, and PCIe). Network connected personal devices – phones, tablets, and laptops – are increasingly being used to manage and interact with these systems. Though conceptually air gapped, the systems are intermittently connected within operational installations to provide mission parameters, or to effect maintenance and upgrades. Though network boundary protections are often in place during these activities, there are many threat vectors that circumvent these protections and airgaps in general; these include unintended network connections, insiders, supply chain interdiction, zero-day exploits, and persistent implants (Kushner 2013). Consequently, we seek embedded devices that may disrupt the attack surface and harden existing systems.

In a previous paper (Dahlstrom et al. 2022), we described a fully automated process that builds a custom hardware traffic validator from a formal language *grammar* describing the message header and payload format. The validators can be deployed in a variety of network appliances, such as the small footprint *guard* pictured in Figure 1, that combines traffic validation with hardware encryption and encapsulation (Dahlstrom and Taylor Aug 2019, Oct 2019, 2020).

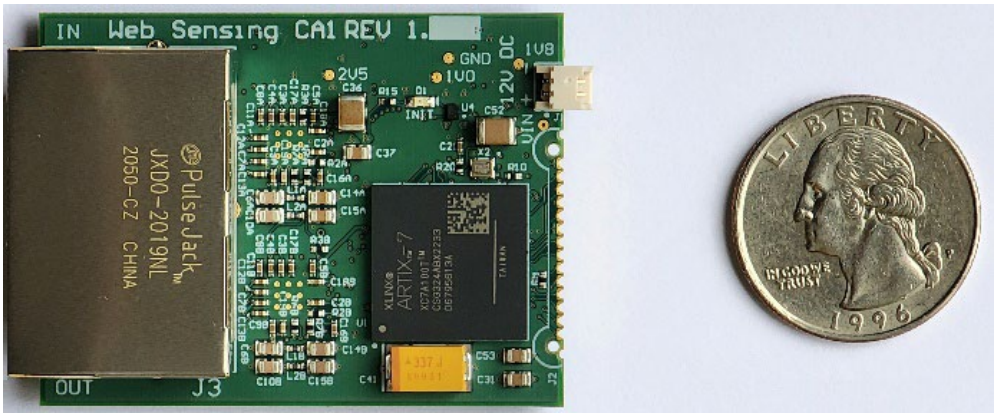


Figure 1: Small Footprint Guard (Size: 2"x1.5"x0.7", Weight: 0.7oz, Power: ~2W)

At their heart each guard consists, by design, of a ***single*** Field Programmable Gate Array (FPGA) lying between industry standard buses and network connections. The guard thereby forms a "bump-in the-wire" with the FPGA acting as a communication bridge. Consequently, the FPGA can monitor and interact with all systems attached to its interfaces, and may act to store, validate, drop, or forward traffic; for example, the guard shown in Figure 1 forms a bridge between two Gigabit Ethernet (GigE) interfaces using the dual connector on the left of the device. Guards offer several key security advantages: All sensitive data -- encryption keys, buses, and algorithms -- is strictly hidden within the security perimeter afforded by the FPGA chip thereby mitigating reverse engineering; No software is present in the device, mitigating persistent software implants and zero-day attacks (Taylor 2018).

2. Sequence Combinators

Traffic validation is, in essence, a *parsing* process in which a string of symbols or values – corresponding to an Ethernet Frame – is analyzed for conformity with respect to a formal language grammar (Grune and Jacobs 2007). Classical *parser generators*, such as Bison (Levine 2009), are tools that automatically construct a software parser implementation from a given grammar. In recent years there has been considerable interest in expanding the expressiveness of parsers, grounded in formal methods, to cope with increasingly complex data formats. Some tools achieve this by leveraging a Domain-Specific Language (DSL) to describe the format (e.g. Levillain 2014, Sommer, Amann and Hall 2016); while others employ *combinators*: primitive elemental parsers with well-defined methods for combining them into more expressive parsers (Hutton 1992, Chifflier and Couprie 2015, Couprie 2015, Bogk and Schöpl 2014, Krishnaswami and Yallop 2019, Bangert and Zeldovich 2019). To achieve this, combinators are often provided in a library that allows different parsers to be constructed by invoking combinator functions using some convenient programming language. For example, the Hammer combinator library (Bratus 2016) allows parsers to be constructed via a C-program.

Unlike this body of research, here we are concerned with what mechanisms admit to the efficient realization of parsers in *hardware* using FPGA's; in particular, if the combinator notion can be used directly to affect an improvement in performance or FPGA resource utilization. Consequently, to validate traffic, the guard shown in Figure 1 implements a standard Left-to-Right Look Ahead (LALR) *parser* that operates on Ethernet frames, one-byte at a time, as they pass through the device. The parser checks that each input byte is valid with respect to a grammar expressed in either Bison Backus-Naur Form (BNF) or Hammer. For example, the following grammar, written in BNF describes, a digit used in parsing JavaScript Object Notation (JSON) (Bray 2017) numbers:

digit : '0' | onenine ;

onenine : '1' | '2' | '3' | '4' | '5' | '6' | '7' | '8' | '9' ;

Reading the grammar: a *digit* is either the terminal symbol '0' in ASCII ***or*** (c.f. '|') is parsed through an alternative rule designated using the non-terminal symbol *onenine*. The *onenine* rule designates nine alternatives, the symbols '1' through '9', explicitly listing all the alternatives. The equivalent Hammer representation is:

H_RULE(zero, h_ch('0'));

H_RULE(digit, h_choice(zero, onenine, NULL));

H_RULE(onenine, h_ch_range(0x31,0x39));

Both these grammars can be transformed into a collection of shift/reduce *parsing tables* that drive a standard push-down automaton that operates with a single stack. Note, however, a subtle conceptual difference between the phrasing: the *h_ch_range() sequence combinator* need simply test that the input is between '1' and '9' rather than include parsing states that test against every potential value in the range. It is this key observation that we exploit here in producing highly compact hardware implementations on FPGA's. To make this concept explicit across both inputs forms, we have extended BNF using a pre-processor – xBNF - that allows the character sequences to be stated explicitly. The equivalent xBNF rules are:

digit : '0' | onenine ;

onenine : ['1' - '9'] ;

3. Automated Parser Construction and Optimization

Figure 2 shows the automation process used to transform a grammar into a hardware parser-plugin that can be deployed on a guard such as that shown in Figure 1. For grammars written in Backus-Naur Form (BNF), the Bison

parser generator tool (Levine 2009) is used to produce a collection of shift/reduce parsing tables, expressed in a machine-readable XML format, describing how the associated push-down automaton should parse the input (leftmost brown path). In Hammer, an equivalent parser can be defined using pre-existing combinators in C linked with the Hammer library (alternate blue path). We have augmented the Hammer library to output the same XML format as Bison. A conversion tool – *xml2h* – is then used to convert the XML parsing tables into a C-header file (pda.h) whose primary content is a large two-dimensional C-array. Rows in the array correspond to *states* in the push-down automaton, while columns designate terminal and non-terminal symbols encountered when reading the input stream. Entries in the array designate *shift* or *reduce* actions applied in each state. Consequently, the C-array provides a complete definition of how the push-down automaton should operate to validate any grammar. The C-array (pda.h) is combined with a generic LALR *automaton* (pda.c) (Aho, Sethi, and Ullman, 1988) and *testbench* code (main.c) to produce a runnable C-program implementing the parser. This software parser is validated using a set of representative test vectors that are successively loaded from files by the testbench (main.c) to ensure that the parser operates correctly.

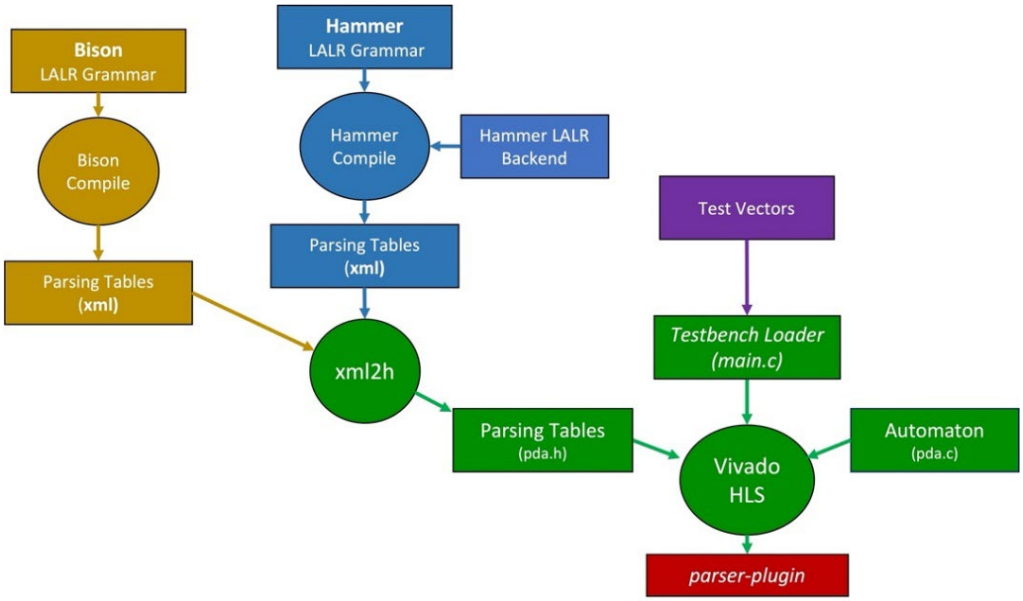


Figure 2: Automated Parser Generation Process

The parser code is carefully constructed to operate on streaming-data and feed directly into the Xilinx High-Level Synthesis (HLS) tools. These produce a hardware circuit block that implements the software parser derived from the pda.[ch] files. The HLS hardware-software co-simulation tool is then used to validate the hardware parser. It uses the same software testbench and test vectors, used to validate the software version of the parser, to ensure the hardware version operates correctly. The validated hardware is subsequently output as a *parser-plugin* component that can be loaded directly into a network appliance, such as that shown in Figure 1. The plugin is organized to connect with the communication interfaces on the board through industry standard AXI-streams.

Unfortunately, though a conventional two-dimensional C-array (pda.h) is a convenient conceptual framework for organizing the automation process, it is impractical since it contains many states that cannot be reached in practice. Consequently, a highly optimized alternative representation is used that removes much of the empty structure from the array. These *compiler*-style optimizations follow similar techniques to those employed internally by Bison and are described in detail online (Popuri, 20006) and in our previous work (Dahlstrom et al, 2022).

4. Combinator Optimization

We have developed an optimization strategy called *the Sequence Combinator* that leverages the concept described in Section 2, namely that when testing ranges of symbols, direct range calculations are used instead of parsing table entries. The optimization is inserted after parsing tables are generated in Figure 1, as shown in Figure 3 for the Bison path, and thus applies to both BNF and Hammer. The optimization digests parsing tables in XML format (gmr.xml), after compiler-style optimizations have been performed, and produces a compressed set of parsing tables, also in XML format, which are then translated into C format using *xml2h*.

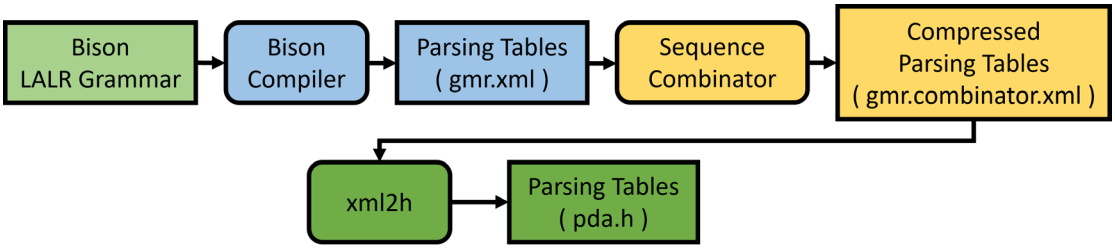


Figure 3: Sequence Combinator Optimization

The *Sequence Combinator* operates by combining sequences of two or more consecutive terminal symbols, $t \in \mathbb{R}^{2:N}$, into a single 'combinator' token, $t_c \in \mathbb{R}^1$, that represents the entire range of token values of t. The subtleties of the transformation are best described using a simple example of a grammar, G, defined in BNF format that accepts any lowercase letter ['a'-'z'] or the string, "foo":

G : letter | string ;

letter : 'a' | 'b' | … | 'z' ; // (26 rules)

string : 'f' 'o' 'o' ;

Notice that the resulting parser will need to distinguish when parsing an 'f' or 'o' which of the letter or string rules are to be reduced by virtue of context.

An un-optimized parser will compile this grammar into a set of parser tables requiring: 2 rules for the non-terminal symbol "G", one rule for each character in the set ['a'-'z'] (26 total), and one rule for the sequence of bytes in "foo". The *Sequence Combinator* optimizes this code by combining consecutive sequences of terminal tokens having only one dependent rule (unlike the terminals 'f' or 'o', which exist on the right-hand-side of both non-terminal symbols "letter" and "string"), such that the new logic becomes:

G : letter | string ;

letter : $t_{a,e}$ | 'f' | $t_{g,n}$ | 'o' | $t_{p,z}$; // (5 rules)

string : 'f' 'o' 'o' ;

where $t_{a,e}$, $t_{g,n}$, $t_{p,z} \in \mathbb{R}^1$and represent the tokens for ranges ['a'-'e'], ['g'-'n'], and ['p'-'z'], respectively. This optimization reduces the number of rules needed to represent the 'letter' symbol from 26 rules down to 5 rules and the total number of states in the PDA is reduced from 34 states down to 13. The optimization algorithm is described in pseudocode in Figure 4.

INPUT: gmr.xml file (e.g. PDA parsing tables)
OUTPUT: gmr.combinator.xml file
INITIALIZATION:
0. read file gmr.xml
BEGIN
1. discover symbol ranges
2. find dependent rules
3. find dependent states
4. add new symbols
5. remove unused terminals
6. renumber token-value of symbols
7. find old rules
8. remove old rules
9. create new range rules
10. renumber rules
11. renumber the states
12. write file 'gmr.combinator.xml'
END

Figure 4: Sequence Combinator Optimization Algorithm

The optimization algorithm, shown in Figure 4, is initialized (Step 0) by reading in the Parsing Tables defined in XML format. This file contains the rules, symbols, and states required to implement the grammar parser using a PDA. Step 1 iterates over all rules in the grammar and discovers sets of continuous symbols that map to a single terminal symbol (e.g., 'a'→ "letter", 'b'→ "letter", etc.). Steps 2 and 3 create lists of rules and states that depend on symbols within a given range which will be compressed. For example, this stage of the algorithm discovers that the non-terminal symbols "letter" and "string" are both dependent upon the terminal symbols 'f' and 'o'. Therefore, these symbols must be excluded from the sequence compression algorithm so that each of the terminal symbols can be represented explicitly in the PDA state machine. Step 4 creates new terminal symbols for the compressed ranges (e.g.['a'-'e'] → $t_{a,e}$). Step 5 removes the unused terminal symbols that have been removed by the compression. Step 6 renumbers the token-values for the reduced set of terminal symbols. Step 7-10 are used to find the old rules, remove them, and then renumber remaining ruleset. Step 11 iterates over all states within the PDA and renumbers the old states and references to old rules – creating the finalized grammar object. Lastly, Step 12 writes the new grammar object to file, 'gmr.combinator.xml' where it can be passed to the xml2h.c script to translate the grammar definitions into a PDA.

An additional modification is required within the PDA implementation (pda.c) to map bytes read from the input to their compressed range token (e.g. 'b' → $t_{a,e}$). This logic could be trivially implemented at runtime by mapping the input byte, B, to a compressed token value if it lies within the min/max bounds (for example: 'a'≤ B ≤ 'e' → $t_{a,e}$) but this would incur additional processing during runtime and limit the bandwidth of the PDA. In our research, we discovered that the mapping of symbols to their respective compressed tokens can be achieved *without any additional computational overhead at runtime* by encoding the symbol mapping into the lookup table at the front end of the PDA. For example, let B represent a byte read from the input and $0 \leq B \leq 255$. The original grammar, G, has 28 terminal characters ('a'-'z', 0, and $end) in the Parser Tables. A lookup table, named "tokens", is used to map B ∈ $\mathbb{R}^{256}$ to a non-terminal symbol, S ∈ $\mathbb{R}^{28}$, which is implemented using an array in C, simply as:

S = tokens[B];

0 = min(tokens)

27 = max(tokens)

The Sequence Combinator compresses grammar, G, into only 7 tokens ({ $t_{a,e}$, $t_{g,n}$, $t_{p,z}$, 'f', 'o', 0, $end}) and the new xml2h script modifies the tokens array to encode the compressed mapping in the following manner:

$S_{a,e}$ = tokens[B];

for B={'a', 'b', 'c', 'd', 'e'}

0 = min(tokens)

6 = max(tokens)

The result of this encoding is that the data read from the input is automatically mapped to the compressed symbol representation, during runtime, without requiring any additional logic or processing overhead.

5. Datatype Optimization

The HLS process shown in Figure 2 is generally performed on standard C code using standard fixed width datatypes. For example, consider the following constant array definition describing the left-hand size of rules in a grammar:

uint16_t lhs[4] = { 1, 11, 2, 13 };

Notice that no value in the array requires more than 4 bits to represent. HLS provides a variable width data specification that makes it possible to use an alternative hardware representation for all data structures used by the push-down automaton. Using this feature, we transform all arrays based on their content and size the data types appropriately:

ap_uint<4> lhs[4] = { 1, 11, 2, 13 };

The result of this transformation is that hardware busses and registers used to represent the *lhs* array in hardware are reduced from 16 bits to 4 bits wide and therefore consume less resources.

6. Resources and Performance

Table 1 reviews the C-structured parsing table data structure sizes prior to HLS, associated with various optimization levels for a full JSON parser and provide insight into the complexity reductions achieved. Notice that these results indicate a dramatic reduction in the size of the structures from a baseline without optimization (-O0), adding compiler-style optimizations that remove sparse areas of the tables (-O1, -O2), and then the subsequent addition of hardware combinators (-O10, -O11, -O12) to each of these baselines. The use of combinators in isolation (-O0 to -O10) reduces tables sizes by 81%; when combined with standard compiler optimizations (-O0 to -12), table sizes are reduced by 95%. Obviously, the Datatype optimization has no impact until synthesis is implemented as it relies on HLS capabilities.

Table 1: JSON Parsing Table Sizes impacted by Combinators

OPT	-O0	-O1	-O2	-O10	-O11	-O12
Terminals	227	227	227	56	56	56
Non-Terminals	48	48	48	48	48	48
Rules	302	302	302	131	131	131
States	346	346	346	175	175	175
Tables (bytes)	192,020	32,552	28,688	37,436	12,716	8,852
% Abs Diff	0.0	0.83	0.85	**0.81**	0.93	**0.95**
% Rel Diff	0.0	0.83	0.12	0.81	0.66	0.3

Tables 2 and 3 shows the actual FPGA resource utilization and average cycles per byte, over a range of representative tests, for two non-trivial parsers used for validating JSON (Bray 2017) and URI (Berners-Lee, Fielding, and Masinter 2005) formats respectively. The first line provides the baseline with no optimization; the second shows with Compiler-style optimizations (-O1 and -O2 in Table 1); the other lines show the cumulative impact of Combinator and Datatype optimizations described in Sections 4 and 5 respectively. Resource counts are for Block RAM (BRAM), Lookup Tables (LUT), Flip Flops (FF), and Digital Signal Processor (DSP) cells. Bandwidth is based on the clock cycles per byte using a 125MHz clock. For comparison, the last line in each table shows the percentage of the Artix 100T resources used with all optimizations applied; This FPGA is used on the small footprint device shown in Figure 1 and has 270 BRAM, 63,400 LUT, 126,800 FF and 240 DSP resources.

Table 2: JSON Parser Resources and Performance

OPT	BRAM	LUT	FF	DSP	Avg Cy/Byte	BW (Mb/Sec)
None	40.5	269	187	1	63	15.9
Compiler	8.5	312	251	1	68	14.7
Combinator	4	305	256	0	57	17.5
Datatype	4	310	218	0	57	17.5
% Artix 100T	***1.4%***	***0.5%***	***0.2%***	-		

Table 3: URI Parser Resources and Performance

OPT	BRAM	LUT	FF	DSP	Avg Cy/Byte	BW (Mb/Sec)
None	7.5	3066	2230	1	58	17.2
Compiler	2	3862	2322	1	63	15.9
Combinator	1.5	3842	2315	0	53	18.9
Datatype	1.5	1284	726	0	53	18.9
% Artix 100T	***0.5%***	***2%***	***0.6%***	-		

With all the optimizations applied, both parsers fit easily within the ARTIX 100T using less than 5% of the resources across the board. This high degree of compactness allows for considerable latitude in the complexity of grammars that can be handled. Consequently, it is possible to produce guards in two flavours: low-cost standalone validators that employ smaller pin-compatible FPGA's, or sophisticated guards that combine the functionality of validation with encryption and encapsulation.

The two parsers show a marked difference in the way HLS allocates resources. The optimizations produce a 90% reduction in BRAM resources on the JSON parser, and an 80% reduction on the URI parser. However, these are lower bounds since block RAM is allocated in slabs and the utilization in each slab cannot be easily determined. In the JSON parser, BRAM allocation leaves little residual logic to be realized with LUT and FF resources, consequently we see only marginal improvements in these areas. However, when substantial levels of these resources are used, as is the case with the URI parser, we see a 67% reduction in LUT and a 69% reduction in FF usage.

In both parsers, bandwidth improves by approximately 10% but remains at less than 20Mb/sec – 50 times slower than Gigabit line rate – however, this is sufficient for real-time analysis of MIL-STD-1553 and CAN bus traffic. Recall that parsing is an inherently sequential activity that considers one byte at a time. Consequently, for normal, good traffic every byte in a payload must be parsed from beginning to end to determine that it is valid. Invalid packets are an outlier in terms of performance measurement and were consequently not considered in the test corpus. The JSON and URI grammars are relatively complex compared to tactical traffic protocols and buffering is capable of handling bursty traffic at higher rates, however, performance rather than resource related optimizations must be devised to improve bandwidth further.

The performance figures presented here represent a second level of space optimization over the initial prototype described in a previous paper and produced unexpectedly compact implementations. In part this is appears to be because all the core work in parsing input is conducted at the leaves of the parsing tree where terminal symbols are resolved; consequently, the sequence combinator is frequently applied with considerable impact.

7. In Conclusion

The guard technology described here occupies a middle ground between the fully air-gapped and fully connected embedded systems: It trades a marginal increase in risk for the opportunity to observe, optimize, and interact with networked systems remotely through a guard. This paper has discussed improved automated processes that yield highly compact, custom hardware parsers directly from formal language grammars. In the extreme, this allows every byte of every message – including payload data -- to be inspected and validated against a formal specification.

Obviously, parsing is inherently a byte-by-byte sequential process that drives detection through a potentially large state-machine – consequently, it is not surprising that bandwidth is limited by the length of input and the complexity of the grammar. Many additional sources of bandwidth enhancement exist that can be expected to improve the technology further as it matures: running parsers at higher clock rates; using extra resources to parse multiple frames or components of frames concurrently; widening input and output data streams; using data-flow optimizations within the parsing pipeline etc. Moreover, the technology need not be used in isolation, or, in an all-or-nothing effort to validate every byte: only a subset of message traffic considered "unsafe" might be considered in conjunction with other, higher-performance tests. For example, by quickly checking an IP address or port, the decision might be made to perform whole packet analysis on only a subset of traffic from specific locations.

The low resource requirements achieved here offer not only the ability to produce further optimizations, but also to include other algorithms concurrently with parsing. For example, they have already been combined with AES encryption and decryption blocks, as well as IPSec packet encapsulation techniques, both of which are able to operate at GigE line rates independently.

Distribution

This research is supported by the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views of the Department of Defense or the U.S. Government. The paper is released with the following distribution determination: **Distribution A: (Approved for Public Release, Distribution Unlimited).**

References

Aho, A.V., Sethi, R., Ullman,. J.D. (1988), *Compilers*, Addison Wesley.

Bangert, J. and Zeldovich, N. (2014), "Nail: A Practical Interface Generator for Data Formats", Proceedings of the 2014 IEEE Security and Privacy Workshops, May 2014, pp 158-166.

Berners-Lee, T. Fielding, R., and Masinter, L. (2005), "Uniform Resource Identifier (URI): Generic Syntax", IETF RFC 3986, January 2005.

Bogk, A. and Schöpl M. (2014), "The Pitfalls of Protocol Design: Attempting to Write a Formally Verified PDF Parser," *2014 IEEE Security and Privacy Workshops*, San Jose, CA, USA, 2014, pp. 198-203.

Bray, T. (Ed.), (2017), "The Javascript Object Notation (JSON) Data Interchange Format", IETF RFC 8259 December 2017.

Bratus, S. et al., (2016), "Implementing a vertically hardened DNP3 control stack for power applications", ICSS'16 ACM Proceedings of the 2nd Annual Industrial Control System Security Workshop, pp 1-9. (c.f. https://gitlab.special-circumstanc.es/hammer/hammer).

Chifflier, P. and Couprie, G. (2017), "Writing parsers like it is 2017", 2017 IEEE Symposium on Security and Privacy Workshops (SPW), San Jose, May 2017, pp 80-92.

Couprie, G. (2015), "Nom, A byte oriented, streaming, zero copy, parser combinators library in rust", In 2015 IEEE Symposium on Security and Privacy Workshops, SPW 2015, San Jose, CA, USA, May 21-22, 2015, pp 142–148.

Dahlstrom, J., Guzman, B., Baker, El, and Taylor, S. (2022), "Automatic Construction of Hardware Traffic Validators", European Conference on Cyber Warfare and Security (ECCWS22), June 2022, pp 60-69.

Dahlstrom, J., and Taylor, S. (Aug 2019), "System-on-Chip Data Security Appliance and Methods of Operating the Same", U.S. Patent 10,389,817.

Dahlstrom, J., and Taylor, S. (Oct 2019), "Endpoints for Performing Distributed Sensing and Control and Methods of Operating the Same", US Patent 10,440,121.

Dahlstrom, J., and Taylor, S. (2020), "System-On-Chip Data Security Appliance Encryption Device and Methods of Operating the Same" U.S. Patent 10,616,344.

Grune, D. and Jacobs (2007), C.J.H., "Parsing Techniques – A Practical Guide", Springer Monographs in Computer Science, 2007.

Hutton, G. (1992), "Higher-Order Functions for Parsing". Journal of Functional Programming 2, 3 (1992), pp 323-343.

Krishnaswami, N.R. and Yallop, J. (2019), "A typed, algebraic approach to parsing", Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, June 2019, pp 379-393.

Kushner, D. (2013), "The Real Story of stuxnet", IEEE Spectrum, vol 50, no. 3, Mar. 2013, pp 48-53.

Levillain, O. (2014), "Parsifal: A Pragmatic Solution to the Binary Parsing Problem." In 35. IEEE Security and Privacy Workshops, SPW (LangSec) 2014, San Jose, CA, USA, pages 191–197, May 2014, pp 191-197.

Levine, J. (2009), "Flex & Bison", O'Reilly Media, Inc. ISBN: 978-0-596-15597-1, 25 July, 2009.

Popuri, S.K. (2006), "Understanding C parsers generated by GNU Bison", Sept 2006 (cs.uic.edu/~spopuri/cparser.htm).

Sommer, R., Amann, J., and Hall, S. (2016), "Spicy: a unified deep packet inspection framework for safely dissecting all your data. In Proceedings of the 32nd Annual Conference on Computer Security Applications, ACM 2016, Los Angeles, CA, USA, Dec. 2016, pp 558-569.

Taylor, S. (2018), "Protecting Embedded Systems from Zero-Day Attacks", IEEE NAECON 2018, July 23-26, pp 165-168.

A Strategic Path for Digital Transformation in Cyber Warfare for African Militaries

Mphahlela Thaba[1] and Jabu Mtsweni[1,2]
[1]Council for Scientific and Industrial Research, Pretoria, South Africa
[2]Stellenbosch University, Security Institute for Governance and Leadership in Africa, Faculty of Military Science, Saldanha, South Africa

jthaba@csir.co.za
mtswenij@gmail.com

Abstract: Digital disruption has changed the battlefield and increased its complexity for the war fighter. The modern battlefield continues to increase this complexity, due to the evolution of components that constitute military capability. The technologies, processes and the users are such components. The modern battlefield relies on advanced technologies tapping on high connectivity, are more lethal, precise, and autonomous. Due to this evolution, areas once thought to be safe from conventional attacks are increasingly becoming vulnerable. This evolution of technology and shorter development curves have also increased the prominence of the cyberspace, as a domain of war. However, many militaries, especially in Africa are still operating legacy systems and struggling with modernizing their systems to take advantage of the digital evolution. This paper, therefore, uses a systematic literature review and benchmarking focusing on selected super cyber power nations' indices to propose a strategic path for African militaries to drive digital transformation in their operational environments. The roadmap is proposed to stimulate the establishment and enhancement of African militaries' cyber warfighting capabilities in the digital age. The objectives of this digital transformation path include establishing a digital backbone, where all the sensors, effectors and the deciders are plugged to share information and intelligence.

Keywords: Digital transformation, Cyber warfare, Cyber operations, Cyber defence, Cyber attacks, Africa battlespace

1. Introduction

Tom Abke argues that the development of the first computers in the 1950's was also beneficial in military operations, in applications such as code-breaking, and logistics (Abke, 2022). He further states that towards the end of the 1980s, beginning of the 1990s, militaries began to adopt more advanced digital technologies, such as Global Positioning System (GPS) and military satellites. These developments significantly improved navigation and intelligence gathering in military operations (Abke, 2022).

Consequently, in the early 1990's many militaries were involved in what was then believed to be an information war, using integrated technologies that gave access to information for the war fighter (Dombrowski & Ross, 2008). It is also during this period that many nations were looking at making their militaries more lethal, and survivable, which included ways of decreasing boots on the ground. The era succeeding this, to the current digital battlefield, has also seen an increase in the use of modern technologies, which are highly connected, advanced, lethal, and intelligent.

The prominence of unmanned vehicle systems, and their evolution into autonomy has made the battlespace even more complex. Today's wars are fought amongst communities, the technologies required for war are dependent on systems that communities require for their livelihoods, and vice versa. This also makes it difficult to determine the boundaries of war, and more exacerbated by the prominence of cyberspace as a domain of war. Many militaries are therefore leveraging the capabilities of the latest technologies, to optimize skills and techniques of the soldiers (Billinga, et al., 2021).

Nevertheless, research reports suggest that African militaries are lagging in digitally transforming their capabilities to be effectively prepared for the modern battlefield in the cyberspace. Therefore, the main objective in this research paper is to investigate the digital transformation journeys of military superpowers to dominate cyberspace. This is done to develop a strategic path that African militaries could adopt in digitizing their capabilities in a systematic manner.

The rest of this paper is structured as follows: Section 2, provides background literature on the modern battlespace, demonstrating how digital transformation is giving dominance to the cyberspace, and the African battlespace is also discussed. In Section 3, the research approach for this study is presented as well as limitations. The concept of digital transformation is presented in Section 4. The benchmarking results based on selected superpowers are presented in Section 5. The main contribution of this paper is presented in Section 6 providing a strategic roadmap for African militaries to digitally transform their capabilities as well as the workforce for cyber warfare. The paper is concluded with recommendations in Section 7.

2. The Modern Battlespace

The future battlespace will be dominated by the evolution related to technologies, the advancing soldier, and the agile doctrine guided by these changes. As we advance towards the fifth industrial revolution, military operations will be dominated by even more advanced weapon systems, which are vastly connected and able to share data and intelligence, leading to better situational awareness. As these weapon systems improve, their sustainability on the battlefield also improves, with new energy sources being developed for them, and their ability to seamlessly relief each other or support in battle.

Figure 1 illustrates the connectivity of the modern battlespace, giving dominance to the cyberspace.

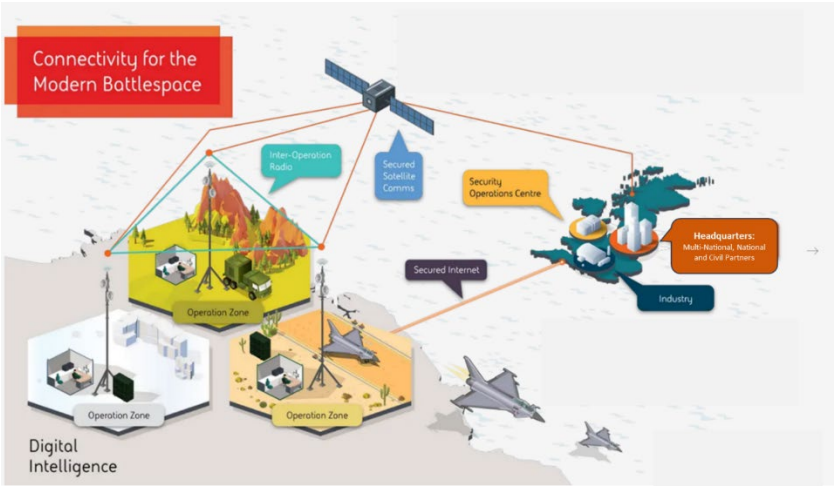


Figure 1: Connectivity of the Modern Battlespace (Adapted)[1]

The evolving technology necessitates changes in how militaries approach warfare. This phenomenon is already manifesting itself in the recent wars, including Russia – Ukraine conflict, where it has been reported that the Wagner group was engaged in the conflict using advanced digital tools (Williamson, 2009). The same has been reported on the African Continent, including alleged involvement of these type of groups in the Mozambique conflict. Global and regional powers also recognize the fact that no one nation will be able to fight the modern battlefield on their own, and many strategize around coalitions, and alliances. The increasing expansion of terrorist and insurgent groups across the world will remain a challenge in the future battlespace because of unrestricted access to digital technologies.

The availability of technologies, infiltrated globally, makes the pace of military offensive fast, and leaves little room for those attacked to respond. Adding to these challenges is cyberspace, which integrates all domains, and is borderless (McGuffin & Mitchell, 2014). Warfare in this domain will continue to be complex and will require militaries to have a complete shift of paradigm in how to plan and develop capabilities for this domain. Battlespace is at a point of no return, where technologies are relied upon for effective and efficient operations (Correia, 2019).

2.1 The African Battlespace

The days of only advanced nations having access to advanced technologies are over. Multinational companies, renowned for developing innovative military technologies, are rapidly increasing their footprint in Africa. Many African militaries accept that to remain ready and capable of fulfilling their mandates, they will have to evolve with the operational environment. The same remain with the increasing terrorists and insurgent groups on the African continent, with links to the base groupings in countries in the middle east and Europe.

The evolution of technology even in the African battlespace, implies the prominence of the cyberspace, and calls for African militaries to plane and develop capabilities that will exploit this complex domain. Whether it is through technology such as intelligent planning tools, digital twins, virtual and augmented reality, or combating threats in new domains, the dividing line between the cyber and the physical world is beginning to blur.

[1] https://www.baesystems.com/en/digital/blog/introducing-trinity-the-future-of-connectivity-for-the-modern-battlespace#

The African cyberspace has matured over time and has in recent times seen emergence of cyber warfare related activities used in conflicts. Recently in the Mozambique conflict, a Yemen Cyber Army claimed responsibilities for cyber-attacks on several websites from various government entities and ministry of defence in Mozambique (All Africa, 2022). Following the reports that Boko Haram hacked the personnel records database of Nigeria's secret service (Baken, 2013), the Nigerian Chief of the Army ordered the creation of the Cyber Warfare Command as well as the Cyber Warfare School (The Nigerian Army, 2022). On 24 January 2022, in Mali, reports surfaced of information campaigns designed to launch cyber-attacks on humanitarian relief organizations2. The infiltration of technologies such as the use of drones as weapons in the Horn of Africa, Sahel and Mozambique by extremist armed groups, state actors or their proxies also signals the emerging trend of the hybrid war, and the exploitation of the cyberspace. (Thaba and Mtsweni, 2023).

3. Research Approach

The evolution of technology has prompted the authors to think about how military operations are affected, especially in Africa. The world over, militaries continue to review their capabilities, in line with the changing battlespace. This paper investigates how some militaries plan to transition their capabilities for the future battlespace. The nations studied in this paper have strategized digital transformation to help their militaries transform for effective operations in the future battlespace. The authors systematically review these open literatures and draw conclusions on how African militaries can use digital transformation to prepare their military capabilities for the modern battlespace and prepare for the cyberspace operations. The authors will also review how the transition could be measured to ensure success of implementation.

4. Digital Transformation

As acknowledged by various militaries, especially those of advanced nations, digital transformation is critical to achieving high levels of operational efficiency and effectiveness. Most of these nations have undertaken the process and invested lots of money in transforming their militaries for future battles. The general objectives of digital transformation, will be to enable the militaries to:

- Make informed, better, and pro-active operational decisions.
- Have superiority of information over adversaries.
- Better Situational Awareness allows them to adapt more quickly and effectively than adversaries.
- Improve competitiveness and operational effectiveness.
- Efficient use of resources by diversifying capabilities that take advantage of the evolving battlespace.
- Plan, prepare and to operate in cyberspace.

One of the driving factors to transforming military capabilities is the availability of data that makes them more intelligent and lethal, the effects required remain relevant with the evolution (Office of the National Director of Intelligence, 2021). The digital transformation addresses the ability to collect, process and transmit data relevant to operations. Migrating from legacy systems in this regard remains a great challenge. This stems from an already existing challenge of integration and interoperability of current systems. Digital transformation, if achieved, would address these challenges. Where new advanced technologies are inserted, the drive to ensure they are interoperable should be one of the challenges to be addressed by digital transformation. The digital backbone must be versatile to be able to handle evolving technologies (Abke, 2022).

5. Benchmarking

In this section, we systematically selected countries that have considered, planned, or are already on their journey to digital transformation to benchmark against African militaries. According to Scott (2011), benchmarking is a systematic approach of comparing processes, outcomes, products, services, and strategies of organizations, nations, or governments for improvements, lessons, and/or best practices. For this research study, we selected five (5) countries based on the indicators as shown in the Table 1 to benchmark against.

The National Cyber Power Index (NCPI) is used to measure the cyber intent and capability of nations in the cyber space and is limited to 30 countries, mostly first world countries (Voo, et al., 2022). The indices of the top countries are based on publicly available data focusing on 29 capability indicators across eight (8) cyber objectives. This simply means a nation is measured based on the intent to pursue multiple objectives using cyber

[2] https://www.icrc.org/en/document/cyber-attack-icrc-what-we-know

means and has the required cyber capabilities to pursue and achieve the setout objectives. In the benchmarking exercise, we only focus on the superior and/or mature capabilities for each of the selected countries.

Table 1: Selection of countries for digital transformation and capability benchmarking

Selected Country	Digital Transformation Strategy Status	Digital Military Capabilities (Current or Envisioned)	Cyber Warfare Capability, Intent & Rank
United States of America	US Army Digital Transformation Strategy (2035 - Aimpoint)	Organic Digital Workforce Data-driven Army Cloud Native Adoption Strengthen collaboration and interoperability of data, software, and systems	-Ranked #1 -Cyber Offense -Information Control -Cyber Norms -Cyber intelligence
Republic of China	China's Military Modernization (2050)	Informatization AI and Advance Robotics Biotechnology Hypersonic and Energy Weapons	-Ranked #2 -Cyber Surveillance -Cyber Financials -Information Control -Cyber Commercials
United Kingdom	UK Army Digital Transformation Strategy	Artificial Intelligence One Defence Upskilling Data-driven military capabilities Digital Backbone	- Ranked #4 -Cyber Defense -Cyber Norms -Cyber Intelligence
Russia	Unknown	Unmanned ground, underwater, and air systems Artificial Intelligence in Weapon Systems	- Ranked #3 - Information Control - Cyber Defense - Cyber Intelligence - Cyber Commerce
Egypt	Digital Egypt (Vision 2030)	Unmanned Aerial Vehicles Surveillance	- Ranked #24 - Cyber Defence - Cyber Surveillance - Cyber Norms

5.1 United States of America

The US leads in various cyber warfare capabilities and has a digital transformation strategy (2035) in place for the US military (Office of the Army Chief Information Officer, 2021) that seeks to drive organic digital military workforce, and data-driven military operations. The strategy posits the modernization of the US Army's underlying network and computer infrastructure, which is essential to their success in the modern battlefield. The US digital transformation represents a shift in operations and culture that fundamentally changes how can the US defense delivers value through the adoption of advanced technologies such as cloud, data, and artificial intelligence (AI) (Dombrowski & Ross, 2008).

In terms of cyber warfare capabilities, the US is also ranked in the first position with superior cyber destructive capabilities, including control and manipulate the information environment. The US also plays a critical and leading role in the development of cybersecurity norms and technical standards across different forums including the United Nations. The US military also has a strong and superior cyber intelligence gather capability to deal with foreign and local threats including cyber terrorism. On the limitations, the US comes second on cyber technology competence, financial gain through cyber means, and cyber surveillance.

5.2 Republic of China

The Republic of China believes that emerging technologies will shape and increase the speed of warfare, meaning future military success will require forces that are "mechanized, informatized, and intelligentized"

(Horowitz & Kahn, 2021). As a result, China have achieved significant progress in their Digital Transformation Strategies, and strategic plans for Emerging and Disruptive Technologies such as Data and Artificial Intelligence. For example, China's 2016 document "China's National Defence in the New Era" aims to make significant progress in achieving the "Informatization" goal representing digital transformation by 2020; and implement the "Intelligentisation" concept by 2035-2050." (Dombrowski & Ross, 2008).

In the Cyber Power Index (CPI), the Republic of China is ranked #2, respectively. According to the CPI, China has a very strong cyber surveillance capabilities to deal with domestic and foreign cyber and traditional threats. Moreover, China is way ahead in using cyber means for amassing and protecting its financial power as well as localization the cyber industry working with public and private partners. From the Chinese case study, it is also clear that there is a strong link between the military and industry partners, including R&D institutions. This collaboration is interesting and indicates that future wars will be fought by the military collaborating with civilian stakeholders.

5.3 United Kingdom

The United Kingdom (UK) has one of the comprehensive military digital strategies amongst the superpowers. According to the United Kingdom's (UK) strategy (UK Army, 2023). The military capability elements in the physical world must be created in the digital world. Then the digitalization of relationships will be reflected in the virtual domain. The strategy continues to note that there are operational behaviors in which physical assets are utilized, and that these activities are carried out for the execution of tasks and duties. They include situational awareness (including evaluation), planning, decision-making, command, and administrative activities. These occur via processes that are well-defined in various forms of regulations. The outcomes of these activities create outputs or actions in the main mission areas of Force Employment, Force Development, and Force Sustainment.

Nevertheless, the UK lags in terms of digital and military capabilities against the superpowers being ranked#4 in the CPI having a strong cyber defence, are involved in the development of cyber norms and technical standards and have demonstrated the intent and capability of cyber intelligence gathering.

5.4 Russia

According to the benchmarking exercise, the Federation of Russia was found to be the only superpower with no publicly available digital transformation strategy for its military. Nevertheless, based on public documents, Russia has demonstrated digital capabilities for the modern battlefield (Lewis, 2022), and these include unmanned ground and underwater systems, AI in weapon systems as well as information control.

Russia is ranked #3 in the CPI. They exhibit a superior intent and capability in the control and manipulation of information environment, their cyber defence capabilities have been demonstrated during the Russia-Ukraine conflict and they have also demonstrated intent of building home-grown cyber capabilities without much defence on foreign countries.

5.5 Egypt

Egypt has a digital transformation strategy that cuts across the different national domains underpinned by vision 2030. Specifically on the cyber warfare, Egypt is ranked #24 out of 30 countries due to their developing cyber defence and surveillance capabilities. In addition, Egypt appears to be playing a participatory role in the development of cyber norms and technical standards. It is worth noting that Egypt is the only African country that is included in the rankings based on publicly available information.

It is evident that superpowers are gearing ahead in their digital transformation journeys to prepare and participate in the modern battlefield. In Africa, it is also clear that there is still a lot of work to be done to catch up in the 4th industrial revolution, and as such the lessons from other nations could be valuable in provided a roadmap for digital transformation in Africa, and this is proposed in the next section.

6. A Strategic Path to Digital Transformation for African Militaries

Digital transformation and cyber warfare capabilities are interconnected; hence this paper proposes that digital transformation to enable militaries to exploit the cyberspace and use its integrating factor to increase battle dominance. The cyberspace is imposed in military operations due to the evolving technological landscape. This implies that capabilities to operate in this domain are critical. There are both opportunities and challenges due to the adopting of digital technologies, and the dominance of the cyberspace for militaries. The challenges include the security vulnerabilities that come with digital technologies, as these create more entry points that can be possibly created by adversaries (World Economic Forum, 2022).

A survey conducted by Jang-Jaccard et al in 2014 established that as digital systems become more pervasive, the potential impact and sophistication of cyber warfare activities also increase (Jang-Jaccard & Nepal, 2014). This is true for and affects military operations. In this evolving operational environment, militaries must establish capabilities to enable operations in the cyberspace. These capabilities must be able to secure and exploit the cyberspace. The details of the required cyber warfare capabilities are addressed in the framework proposed in Thaba and Mtsweni (2023).

6.1 Capability Definition

Figure 2 overleaf depicts the proposed framework or strategic path towards digital transformation, with the specific focus on how African militaries can implement this. The path uses the Capability Lifecycle model (Smit & Oosthuizen, 2011) representing development phases for the military capability, encompassing related activities geared towards assisting African militaries to achieve digitally transformed capabilities. The model defines the capability in terms of POSTEDFIT model (Thaba & Benade, 2014), encompassing an extended view of the people, processes, and technology. In the capability definition phase, militaries must define the context and understand the strategic direction guiding the force's modernization. This phase must answer the question on why digital transformation is required. This phase must culminate in the development of a Digital Transformation Strategy that will enable African militaries to operate effectively especially within the cyberspace. In the context of cyber warfare, the capabilities requiring transformation are related to the ability to secure and exploit the cyberspace, and continuous improvement in operating cyber warfare capabilities as described in the proposed framework by Thaba and Mtsweni (2023).

However, transforming military capabilities, dominated by legacy systems that have been in operation for decades requires careful thought and consideration, because the threats these capabilities were established for still exist.

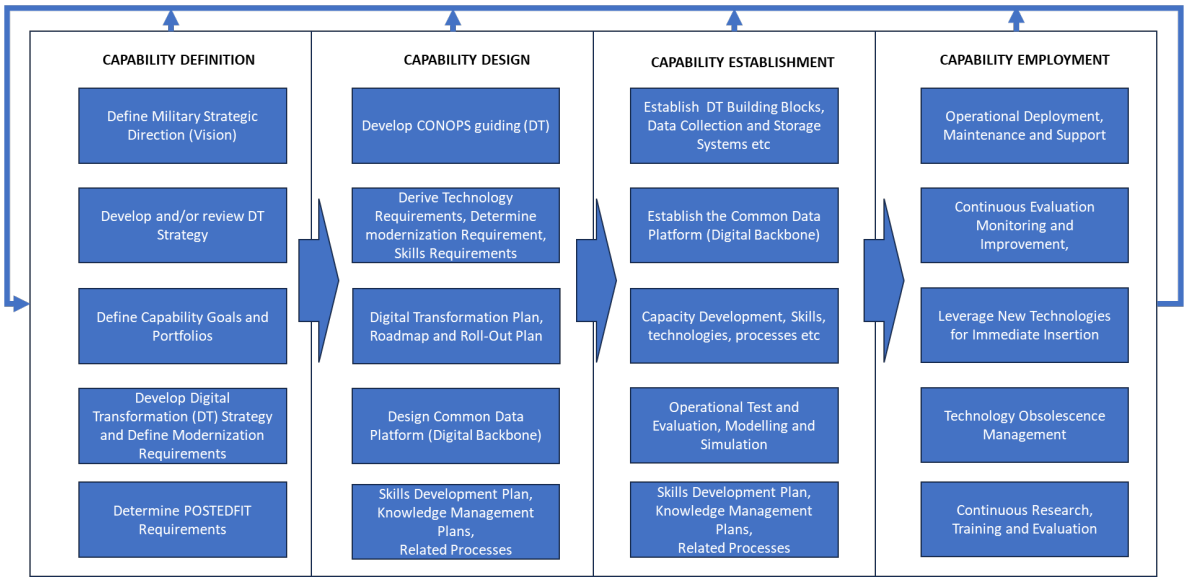


Figure 2: Proposed Path to Digital Transformation for African Militaries

6.2 Capability Design

Digital transformation significantly influences the landscape of cyber warfare by shaping the targets, tactics, and technologies involved in both offensive and defensive cyber operations (Osamo, et al., 2022). During the capability design phase, the efforts must be directed at developing the means to solve the problem as identified in the context in phase 1. During this phase, a concept of operation (CONOPS) must be developed. CONOPS will help derive technology and modernization requirements, leading to the development of the Digital Transformation Plan. This phase must answer the question on how African militaries digitally transform their capabilities to competently operate in the cyberspace. This phase will culminate in the Design of the Common Data Platform, the digital backbone as the critical component required to achieve digital transformation (UK Army, 2023). African military capabilities must be aligned to the current and future challenges and design capabilities for operations in the cyberspace.

6.3 Capability Establishment

This phase involves the establishment of the various building blocks contributing towards a digitally transformed force. As derived from the UK strategy, this must be a secure, singular, modern Digital Backbone (see Figure 3 overleaf) connecting sensors, effectors, and deciders across domains and with partners, driving integration and interoperability (UK Army, 2023). It must be an ecosystem, composed of a combination of people, process, data and technology that will enable friction-free access to our data, connecting sensors in one domain to platforms in other domains, via decisionmakers at the relevant levels in real time (UK Army, 2023).

6.4 Capability Employment

During the capability employment phase, a digitally transformed force will be tasked and deployed to perform various tasks. The phase also includes operational maintenance and support of a digitally transformed force. Technology develops at a fast pace in the modern battlespace; this phase also includes the ability to manage obsolescence and timeously insert technologies available or developed. The phase is concerned with the operational efficiency and effectiveness of the force (Anon., 2018). However, with digital transformation comes the availability of data as a strategic asset and a threat, which are more central to the operational efficiency or effectiveness of the military. However, data availability alone does not create necessary effects; it must be processed and packaged in a way suitable to provide decision support and intelligence to the decision makers.

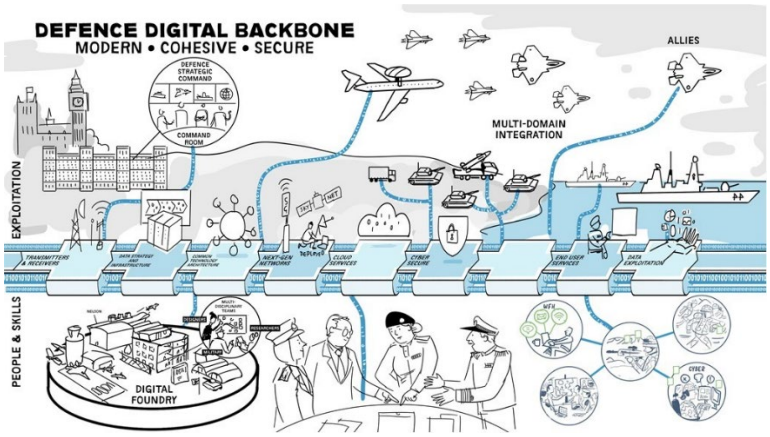


Figure 3: The Digital Backbone (Adapted from (UK Army, 2023))

7. Recommendations and Conclusion

The modern battlespace is characterized by the dominance of cyberspace, which is cutting across all the other traditional domains. This means actions in cyberspace have significant impact in the traditional domains. The boundaryless nature of cyberspace further complicates the battlespace, where the definitions, and boundaries for friend and foe are fluid, and can change any time. The ability to operate in the cyberspace for African militaries is no longer a choice, but a critical necessity. To operate in cyberspace, militaries need the ability to take full advantage of the digital space, by digitally transforming capabilities through digital transformation. In this paper, we propose a capability-driven strategic path for African militaries to transform their capabilities for the modern warfare. The roadmap is supported by an adopted digital maturity model that could aid and support African militaries to systematically develop to equally operate in the modern warfare.

Even for advanced militaries, digital transformation is acknowledged to be a process that will require dedicated effort and resources including time. African militaries must invest significant resources to the effort to be able to achieve this at least in the medium and long term. However critical is the strategic guidance, to give direction and initiation of the process, including availing necessary resources to undertake this process. The strategic guidance role also should be to facilitate this process amongst all the stakeholders, to ensure synergy, and movement towards one common goal. From this paper, the authors have demonstrated the applicability of the Capability Lifecycle management process, which was defined as the guiding phases towards establishing and managing digitally transformed military capabilities.

References

Office of the Army Chief Information Officer, 2021. *Army Digital Transformation Strategy,* s.l.: US Army.

Abke, T., 2022. *Digitalization in the Armed Forces.* [Online] Available at: https://www.linkedin.com/pulse/digitalization-armed-forces-tom-abke/

Anon., 2018. *US DOD.* [Online] Available at: https://irp.fas.org/doddir/dod/jp3_12.pdf

Billinga, D. C., Fordy, G. . R., Friedl, K. . E. & Hasselstrøm, H., 2021. The implications of emerging technology on military human performance research priorities. *Journal of Science and Medicine in Sport,* pp. 947-953.

Correia, J., 2019. Military Capabilities and the Strategic Planning Conundrum. *Security and Defence Quarterly.*

Delloitte, 2018. *Digital Maturity Model Achieving digital maturity to drive growth,* s.l.: s.n.

Dombrowski, P. & Ross, A. L., 2008. The Revolution in Military Affairs, Transformation and the Defence Industry. *Institute for Regional Security,* pp. 13-38.

Global Firepower, 2023. *2023 Military Strength Ranking.* [Online] Available at: https://www.globalfirepower.com/countries-listing.php

Hamilton, J., van der Smissen, S., Ruth, L. & Dailey, L., 2022. *Military procurement in a digital age,* s.l.: Deloitte Center for Government Insights.

Horowitz, M. & Kahn, L., 2021. *DoD's 2021 China Military Power Report: How Advances in AI and Emerging Technologies Will Shape China's Military.* [Online] Available at: https://www.cfr.org/blog/dods-2021-china-military-power-report-how-advances-ai-and-emerging-technologies-will-shape

Jang-Jaccard, J. & Nepal, S., 2014. A survey of emerging threats in cybersecurity. *Journal of Computer and System Sciences.*

Lewis, J. A., 2022. *Cyber War and Ukraine.* [Online] Available at: https://www.csis.org/analysis/cyber-war-and-ukraine

McGuffin, C. & Mitchell, P., 2014. On domains: Cyber and the practice of warfare. *International Journal,* pp. 394-412.

Office of the National Director of Intelligence, 2021. *The Future of the Battlefield,* s.l.: Global Trends.

Osamo, V. C., Azeta, A., Guembe, B. & Misra, S., 2022. The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence.*

Scott, R., 2011. Benchmarking: A literature review. *Academic Excellence Centre for Learning and Development.*

Smit, C. J. & Oosthuizen, R., 2011. *APPLYING SYSTEMS ENGINEERING PRINCIPLES TOWARDS DEVELOPING DEFENCE CAPABILITIES.* s.l., s.n.

Spak, U., 2021. *The common operational picture: A powerful enabler or a cause of severe misunderstanding?.* s.l., ICCRTS.

Thaba, M. & Benade, S., 2014. *Aligning force planning and systems acquisition.* Somerset West, Western Cape, In Proceedings of EMEASEC.

Thaba, M. & Mtsweni, J., 2023. *Developing Robust Cyber Warfare Capabilities for the African Battlespace.* Athens, Greece, s.n.

UK Army, 2023. *The Army Digital and Data Plan 2023 - 2025,* s.l.: UK Army.

Voo, J., Hemani, I. & Cassidy, D., 2022. *National Cyber Power Index 2022,* Cambridge: Belfer Center for Science and International Affairs, Harvard Kennedy School.

Williamson, S. C., 2009. *From fourth generation warfare to hybrid warfare,* Pennsylvania: United States Army.

World Economic Forum, 2022. *The Global Risks Report 2022,* s.l.: s.n.

Adversarial Camera Patch: An Effective and Robust Physical-World Attack on Object Detectors

Kalibinuer Tiliwalidi[1], Bei Hui[2], Chengyin Hu[1] and Jingjing Ge[3]
[1]School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[2]School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan, China
[3]Unit 31153

202011081727@std.uestc.edu.cn
bhui@uestc.edu.cn
cyhuuestc@gmail.com
G_jingjing2023@hotmail.com

Abstract: Physical adversarial attacks present a novel and growing challenge in cybersecurity, especially for systems reliant on physical inputs for Deep Neural Networks (DNNs), such as those found in Internet of Things (IoT) devices. They are vulnerable to physical adversarial attacks where real-world objects or environments are manipulated to mislead DNNs, thereby threatening the operational integrity and security of IoT devices. The camera-based attacks are one of the most practical adversarial attacks, which are easy to implement and more robust than all the other attack methods, and pose a big threat to the security of IoT. This paper proposes Adversarial Camera Patch (ADCP), a novel approach that employs a single-camera patch to launch robust physical adversarial attacks against object detectors. ADCP optimizes the physical parameters of the camera patch using Particle Swarm Optimization (PSO) to identify the most adversarial configuration. The optimized camera patch is then attached to the lens to generate stealthy and robust adversarial samples physically. The effectiveness of the proposed approach is validated through ablation experiments in a digital environment, with experimental results demonstrating its effectiveness even under worst-case scenarios (minimal width, maximum transparency). Notably, ADCP exhibits higher robustness in both digital and physical domains compared to the baseline. Given the simplicity, robustness, and stealthiness of ADCP, we advocate for attention towards the ADCP framework as it offers a means to achieve streamlined, robust, and stealthy physical attacks. Our adversarial attacks pose new challenges and requirements for cybersecurity.

Keywords: Deep neural network, Camera-based physical attack, Object detector, Effectiveness, Robustness

1. Introduction

Deep Neural Networks (DNNs) are among the most important technologies that assist in accomplishing the goals of computer vision. They excel in tasks such as image classification, object detection, and segmentation, as they can automatically learn and extract meaningful features from images. Despite the superior performance of DNNs in real-world applications, there are serious shortcomings in their robustness in adversarial scenarios. Studies have found that DNNs are easily deceived by adversarial samples maliciously constructed by attackers, resulting in inaccuracies in model prediction (Szegedy et al 2013; Goodfellow et al 2014; Guesmi et al 2023).



(a) AdvCS (b) TTP (c) ADCP (Ours)

Figure 1: Demonstration of ADCP and other camera-based attacks

At present, a significant portion of physical attacks (Hu et al 2022; Wang et al 2021; Suryanto et al 2022) utilizes adversarial patches as perturbations to execute physical attacks against advanced object detectors. Adversarial patches commonly cover a substantial fraction of the target object's area, bolstering the robustness of physical

attacks at the expense of reduced stealthiness. While patch-based physical attacks maintain the semantic integrity of the target object, the conspicuousness of the perturbation remains a challenge. To address this issue, some studies have introduced light-based physical attacks (e.g., lasers (Hu et al 2023), projectors (Gnanasambandam et al 2021), shadow(Zhong et al 2022)). These leverage the transitory nature of illumination to instantaneously project an optimized beam onto the target object's surface, facilitating immediate physical attacks. Differing from patch-based counterparts, light-based methods can manipulate the light source's on-off state, projecting the beam during attacks and extinguishing the source when dormant. This results in superior stealthiness for light-based methods, as the physical perturbation isn't persistently affixed to the target object's surface. While light-based physical attacks enhance stealthiness, they often come at the cost of reduced robustness. Similarly, both light-based and patch-based physical attacks share a commonality: they entail modifications to the target object. To tackle this limitation, certain studies have proposed camera-based physical attacks, such as Adversarial camera stickers (AdvCS) proposed by Li et al (2019) and Translucent Patch (TTP) proposed by Zolfi et al (2021), these attacks are all based on white-box attacks that assume complete knowledge of the target model's architecture and parameters, which is often unrealistic in real-world scenarios. In this study, we investigate black-box camera-based adversarial attacks in a more extreme yet practical physical setting. The focus is on generating robust adversarial examples to attack advanced object detectors while maintaining superior robustness, stealthiness, and affordability for real-world applications. A camera patch perturbation is crafted to fool the model for all instances of a specific object class, while maintaining the detection of untargeted objects. Figure 1 vividly illustrates the deployment contrasts between the proposed approach and the baseline strategy. It is clearly that the model correctly recognizes other objects. The contributions of the study can be outlined as follows:

- Introduce a novel camera-based physical attack (ADCP) based on a black-box for robustly generating adversarial samples to execute attacks on advanced object detectors. ADCP utilized PSO to generate robust adversarial perturbations which assist in surpassing baselines in terms of robustness, stealthiness, and deployment simplicity.
- Provide a model with inexpensive and reachable attack material with effortless implementation, where the expected budget is not exceeding $7, to render readily applicable in real-world scenarios.
- Conducting a meticulous comparison of existing physical attack techniques to determine the distinctive benefits of the approach in relation to its equivalents. The comparisons confirm that ADCP is valuable and outperforms recent camera-based physical attacks where ADCP preserves its stealthy nature.
- The study extends to a comprehensive investigation of the proposed ADCP method, including ablation experiments that confirm the method's consistent performance across diverse color settings. These analyses contribute to a comprehensive understanding of the method's performance and its practical viability.

2. Related Works

Patch-based attacks involve the strategic use of meticulously crafted patches affixed to the target object's surface to deceive advanced DNNs (Chen et al 2019). Recent research has focused on achieving a balance between the patch's stealthiness and its impact on attack efficacy. Thys et al. (2019) used total variation loss to generate smoother adversarial patches that compromise pedestrian target detection systems. Building upon this, Wu et al. (2020) printed optimized adversarial patches onto clothing to diminish their conspicuity to human observers, showcasing effectiveness across white-box and black-box settings. Similarly, Tan et al. (2021) introduced a method involving cartoon-like patches that deceive pedestrian detectors, employing a two-stage training approach to generate natural and rational adversarial patches.

On the other hand, camera-based attacks involve applying an adversarial patch onto a camera lens, capturing a physical sample of the target object and subsequently deceiving DNNs. Distinguished by its nonintrusive character, camera-based attacks confer a stealth advantage compared to patch-based, which necessitate target object modifications. Li et al. (2019) introduced the AdvCS camera-based attack, which utilized lens modification to affix small non-transparent patches, effectively deceiving classifiers. It is an uncommon scenario in practical applications. Addressing these limitations, Zolfi et al. (2021) proposed TTP, a refined camera-based physical attack using translucent patches for improved sample concealment, achieving a 42.27% attack success rate against advanced detectors. Yet, the subtlety of TTP came at the expense of attack robustness. In essence, prevailing camera-based methods confer spatial invisibility without affecting target objects. Nevertheless, AdvCS and TTP's reliance on multiple small patches for lens coverage culminated in operational complexity and

substantial experimental errors, limiting attack performance and feasibility. Contrasting these approaches, ADCP circumvents these challenges through the application of a solitary translucent patch as perturbation.

3. Method

In this study, similar to the explanation provided by Zolfi et al (2021), it is suggested that the attacker possesses direct access to the camera lens, enabling the deployment of meticulously optimized perturbations upon its surface.

3.1 Problem Definition

Considering dataset D comprising clean samples, it is divided into two sets, X and Y, representing the collection of clean samples and the collection of ground truth labels, correspondingly. With f representing the pre-trained model of the object detector, for every X belonging to the dataset D, the object detector $f: X \rightarrow Y$ adeptly prognosticates the label y for the clean sample. Within y, there are three important elements: V_{pos}, denoting the positional intel of the bounding box; V_{obj} , encapsulating the confidence level of the target object; and V_{cls}, conveying the category of the prognosticated entity:

$$y = [V_{pos}, V_{obj}, V_{cls}] = f(X) \quad (1)$$

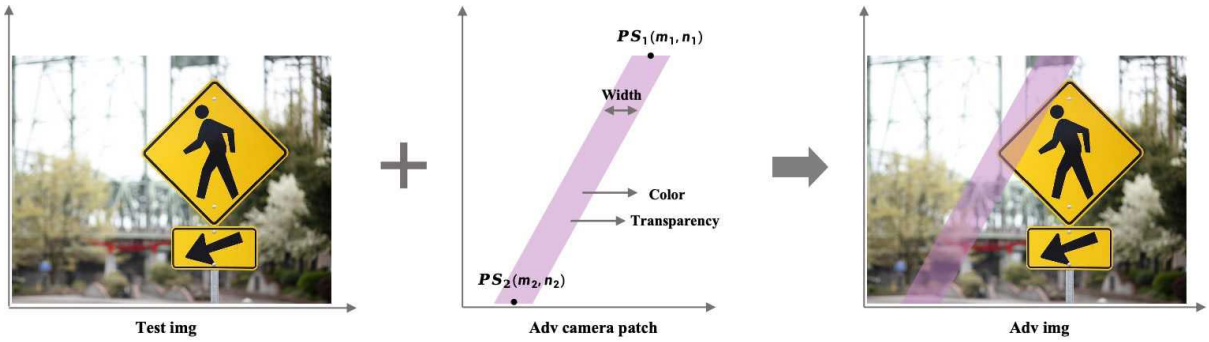


Figure 2: Camera patch modeling

3.2 Camera Patch Modeling

The approach uses a translucent patch affixed to the camera lens to conduct a black-box physical attack. The camera patch is depicted conceptually in Figure 2 and is characterized by four distinct physical parameters: position (PS), color (C), width (W), and transparency (TS).

Position PS: The position parameter is denoted as PS and determined by the coordinates of the two endpoints of a line, pinpointing the precise location of the camera patch within the image. In the experimental setup, we keep n_1 and n_2 as constant values to vary the camera patch's position solely in the horizontal plane of the image.

Color C: In the investigation, color is utilized to encapsulate the visual attributes of the camera patch, with its representation denoted as $\mathrm{C}(\mathrm{r}, \mathrm{g}, \mathrm{b})$, where r, g, and b signify the red, green, and blue channel values of the color, respectively. In digital attack experiments, the choice of color is unconstrained. However, for physical attacks, we exclusively use pink, blue, and green camera patches due to practical limitations. Subsequent experiments demonstrate a minimal correlation between color and attack success rate.

Width W: To effectively encapsulate the scale of the camera patch in the horizontal direction, the parameter width (W) is introduced. As depicted in Figure 2, the width W is designed to measure the horizontal extent of the camera patch. In prescribing the value of this parameter, it is aligned with the width of the image to ensure its adaptability across diverse image dimensions. Specifically, the width W *is defined* within a range spanning from 0.1 to 0.9, with intervals set at 0.2. The chosen value of W signifies the proportion of the camera patch's width to that of the overall image.

Transparency TS: The transparency parameter ranges from 0.1 to 0.9, encapsulating the perceptual visibility of the camera patch. A lower value implies heightened transparency, rendering the camera patch less discernible, while a higher value accentuates its presence, increasing the likelihood of capturing the observer's attention.

3.3 Camera Patch Attack

The vector $\theta = \{PS, C, W, TS\}$ is adopted to succinctly represent the camera patch generated in this approach. The size of θ adheres to the pre-set boundaries defined by the vectors θ_{min} and θ_{max}, where θ_{min} and θ_{max} can be adjusted. Hence, the process of crafting adversarial examples within a digital context by employing predetermined physical parameters can be elucidated as follows:

$$X_{adv} = S(X, \theta) \quad \theta \in (\theta_{min}, \theta_{max}) \tag{2}$$

Where, *S* represents a simple linear fusion of the generated simulation camera patch and the clean sample to obtain the digital adversarial sample X_{adv}.

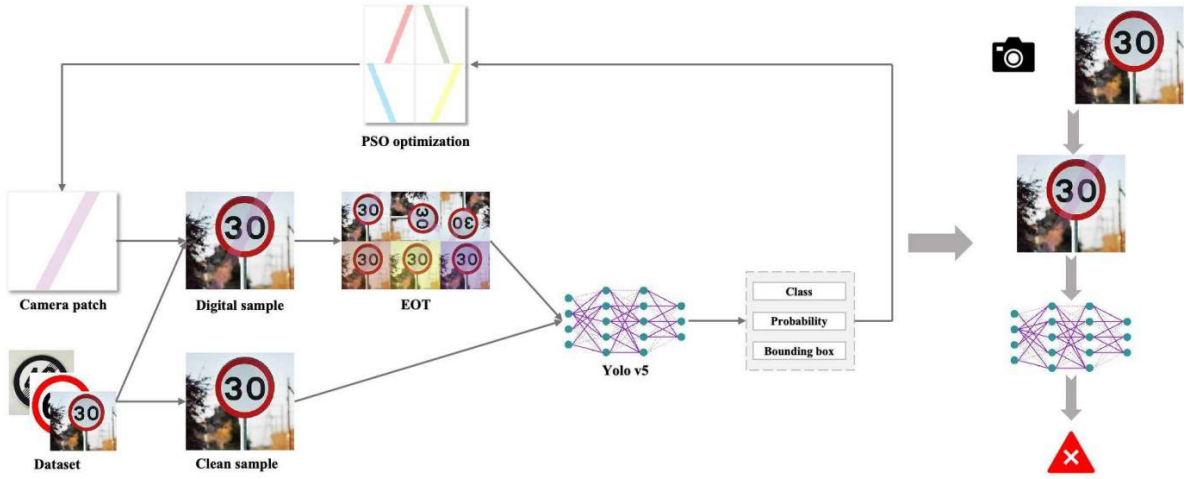


Figure 3: Overview of ADCP attack. The left side shows the optimization process in a digital environment, and the right side indicates that the method generates physical samples in a physical environment. After the robust digital adversarial samples are optimized in the digital environment, perturbations will be deployed in the physical world to generate physical samples

Figure 3 illustrates the attack strategy employed in the study. In the digital field, camera patches with an expansive spectrum of colors are simulated for the execution of the attack. Meanwhile, in the physical scenarios, fixed-color camera patches (pink, green, blue) are opted for the attack. To bridge the experimental gap between simulated and physical samples, the technique of Expectation Over Transformation (EOT) (Athalye et al., 2018) is introduced. Within the context of EOT, a transformation T encompasses the distribution of all possible transformations, serving as the conduit for simulating the transition from the digital domain to physical scenarios. For 2D adversarial examples, the transformation set T includes rotation, variance, dimming, Gaussian noise, image translation, and other relevant alterations. By integrating these transformations, the optimization of adversarial examples that remain adversarial under transformation T is achieved, thus bolstering the efficacy of physical attacks and mitigating the influence of experimental deviations. Through the incorporation of EOT, the impact of experimental discrepancies on physical attacks is successfully mitigated. Consequently, the ultimate representation of the physical sample is encapsulated as follows:

$$X_{adv} = EOT_{t\sim T}(t(S(X, \theta), \theta)) \quad \theta \in (\theta_{min}, \theta_{max}) \tag{3}$$

Within the context of this research, the primary objective centers around the simulation and optimization of the physical parameter θ, a precursor to generating the most potent adversarial camera patch. The pivotal criterion is to produce an adversarial sample X_{adv}, engendered by the physical parameter, that yields successful deception of the object detector. Consequently, the object detector would either be unable to correctly identify the object or be prompted to misidentify it. A distinguishing aspect of this study is its departure from preceding camera-based endeavors (Li et al 2019; Zolfi et al 2021), where the aim is to enhance methodological realism. Specifically, this work operates within the framework of a black-box attack scenario, signifying that access to intricate details such as the target model's network structure is precluded. Instead, only the target class output (V_{cls}) and its associated confidence score (V_{obj}) are accessible. This context precipitates the formulation of a novel approach, wherein V_{obj} is harnessed as the adversarial loss. In light of this, the objective is framed as the optimization of the physical parameters characterizing the camera patch, with the intent of minimizing the adversarial loss. This intricate process seeks to generate adversarial samples that successfully fool the object detector. The crux of this optimization endeavor can be succinctly encapsulated as follows:

$$argmin_{\theta} EOT_{t\sim T}(V_{obj} \leftarrow t(f(X_{adv}), \theta)) \quad (4)$$

To achieve global optimization, the PSO algorithm (Kennedy and Eberhart, 1995), inspired by bird predation behaviors and aimed for rapid convergence and optimal solution discovery, is employed. In the PSO algorithm, each optimization problem's solution is metaphorically represented as a "particle" within a search space. These particles possess distinct fitness values determined by the optimization function and velocity vectors that dictate their navigational trajectories. All particles collectively explore the solution space, following the trajectory of the best particle, known as the "best global particle."

The PSO algorithm begins with the initialization of a set of random particles, representing stochastic solutions. Through iterative cycles, the algorithm navigates the solution space to uncover the best solution. In each iteration, every particle updates its position based on two key values: its individual best solution (denoted as "individual extrema") and the best solution found by all particles (referred to as "global best"). These updates are performed using a formula designed to optimize the particle's movement through the solution space, aligning with the iterative dynamics of the PSO algorithm. The velocity and position of the particles are updated according to the following formula:

$$v_i^{j+1} = \omega v_i^j + c_1 r_1 (P_{i,best}^j - \theta_i^j) + c_2 r_2 (G_{best}^j - \theta_i^j) \quad (5)$$

$$\theta_i^{j+1} = \theta_i^j + v_i^{j+1} \quad (6)$$

where $i \in (1, k)$, k denotes the number of populations. j denotes the current iteration number and ω, c_1, r_1, c_2, r_2 denote the hyperparameters of the PSO algorithm. v is the velocity of the particle and θ is the position of the particle (i.e., the current solution).

Algorithm 1: Pseudocode of ADCP

Input: Input $\boldsymbol{X}$, object detector $\boldsymbol{f}$, Ground truth label $\boldsymbol{Y}$, Max step *Step*, $\boldsymbol{\omega}, \boldsymbol{c_1}, \boldsymbol{r_1}, \boldsymbol{c_2}, \boldsymbol{r_2}$;

Output: A vector of parameters $\boldsymbol{\theta}$;

For each particle $\boldsymbol{i}$:

    Initialize position $\boldsymbol{\theta_i}$ randomly;

    Initialize velocity $\boldsymbol{v_i}$ randomly;

End For

For $\boldsymbol{j} \leftarrow 0$ to *Step*:

  For each particle $\boldsymbol{i}$:

    $\boldsymbol{X_i^j = S(X, \theta_i^j(PS, C, W, TS))}$;

    $[\boldsymbol{V_{pos}, V_{obj}, V_{cls}}] \leftarrow \boldsymbol{f}(\boldsymbol{X_i^j})$;

    Obtain the individual optimal value $\boldsymbol{P_{i,best}^j}$;

    Obtain the global optimal value $\boldsymbol{G_{best}^j}$;

    If $\boldsymbol{f(X_i^j) \rightarrow \emptyset \ or \ V_{cls} \neq Y}$:

      Output $\boldsymbol{\theta = \theta_i^j(PS, C, W, TS)}$;

      Exit ();

    End If

  End For

  For each particle $\boldsymbol{i}$:

    $\boldsymbol{v_i^{j+1} = \omega v_i^j + c_1 r_1 (P_{i,best}^j - \theta_i^j) + c_2 r_2 (G_{best}^j - \theta_i^j)}$;

    $\boldsymbol{\theta_i^{j+1} = \theta_i^j + v_i^{j+1}}$;

  End For

End For

Algorithm 1 outlines the process of the proposed ADCP employing PSO for optimization. ADCP takes several inputs: a clean sample X, the target detector f, ground truth label Y, maximum number of iterations Step, and PSO hyperparameters ω, c_1, r_1, c_2, and r_2, which can be determined by the attacker. The pseudocode details are elucidated in Algorithm 1. Initially, it initializes the initial velocity and position of each particle within the swarm. Subsequently, in each iteration, the camera patch represented by each particle is combined with the clean sample to generate an adversarial sample. The confidence score of the particle, which serves as its fitness value, is then obtained. It's worth noting that a lower confidence score indicates a higher fitness level for that particle. For each particle, if its corresponding adversarial sample induces the model to misidentify or fail to recognize the target, the position information of the particle is recorded as the required physical parameters for the camera patch. Ultimately, in every iteration, the historical optimal solution $\mathrm{P_{best}}$ for each particle and the historical optimal solution $\mathrm{G_{best}}$ for the entire swarm are determined. These two optimal solutions are employed to update the velocity and position information of each particle. The program ultimately outputs the physical parameter θ of the camera patch, which is utilized for subsequent physical attacks.

4. Evaluation

4.1 Experimental Setting

Dataset: In this study, the TT100K dataset (Zhu et al 2016) is chosen as the fundamental source for both model training and attack experiments. This detaset encompasses over 30,000 traffic sign instances extracted from approximately 100,000 images, characterized by a resolution of 2048x2048 pixels and a diverse array of lighting and meteorological conditions. The original dataset is carefully curated to retain only those traffic signs containing no fewer than 100 samples. Consequently, a new dataset emerges, denominated as TT100K-CameraPatch (TT100K-CP), encompassing a total of 10,427 images. Throughout the experiments, the TT100K-CP dataset is divided into training, testing, and validation sets, comprising 7,568, 1,889, and 970 images, respectively. Moreover, for the digital attack experiment, the segregated validation set is designated as the attack dataset, enabling a comprehensive evaluation and validation of the proposed method.

Object detector: In this study, YOLOv5 (Jocher 2020) was selected as the object detector for training the traffic sign detection model for Chinese road signs due to its speed, efficiency, and wide acceptance in the field. To expedite convergence, pre-trained weights from YOLOv5 were used and the model was fine-tuned with the TT100K-CP dataset. This approach accelerated the training process and facilitated model adaptation to the specific dataset, resulting in significant results on the test set, achieving an impressive average accuracy of up to 80%.

Experimental devices: In physical experiments, the experimental devices used included a support frame, camera patches, and a Redmi K40. The support frame was securely installed on the phone to ensure experiment stability and repeatability, enabling the conduct of physical attack experiments in a real-world environment to verify the effectiveness of the proposed approach in practical scenarios. After verification, it was found that the proposed method maintained robustness and feasibility on different camera devices, strengthening the practical application value of the study and making the method potentially applicable to a variety of camera devices.

Evaluation criteria: We embrace the Attack Success Rate (ASR) as the pivotal metric to gauge the efficacy of our proposed methodology. The ASR is defined as follows:

$$\mathrm{ASR(X)} = 1 - \frac{1}{\mathrm{N}}\sum_{\mathrm{i=1}}^{\mathrm{N}} \mathrm{F(label_i)}$$

$$\mathrm{F(label_i)} = \begin{cases} 1 & \mathrm{label_i} \in \mathrm{L_{pre}} \\ 0 & \text{otherwise} \end{cases}$$

where N represents the number of true positive samples detected by the target detector f in the data set D in the case of no attack, and $\mathrm{L_{pre}}$ represents the set of all labels detected in the case of attack. A higher ASR indicates a more effective attack.

Other details: The hyperparameters of PSO as follows: $\omega = 0.9,\ \mathrm{c_1} = 1.6,\ \mathrm{r_1} = 1,\ \mathrm{c_2} = 2.0,\ \mathrm{r_2} = 1$. For all attack experiments, we execute on a single NVIDIA GeForce RTX 2080 Ti GPU.

4.2 Evaluation of Effectiveness

Digital test: In this study, a series of digital experiments is conducted to validate the proposed approach in a simulated environment. Meticulous ablation experiments are carried out to evaluate the performance of the

method in a digital context. These ablation experiments served two main purposes: determining the attack configuration for subsequent physical experiments and analyzing the impact of ADCP under different configurations. Two pivotal factors influencing the effectiveness of the method are identified, namely, the width and transparency of the camera patch. It is observed that excessively small camera patches may fail to yield robust adversarial effects, while overly wide patches could compromise the attack's subtlety. Similarly, highly transparent camera patches may not ensure robust adversarial effects, while those with low transparency might undermine stealth. Thus, the goal is to find a delicate balance between optimizing attack effectiveness and maintaining stealth. The width parameter of the camera patch spans from 0.1 to 0.9 with a step size of 0.2, and the transparency parameter traversed the range of 0.1 to 0.9 with an increment of 0.1. Ablation experiments are conducted across various combinations of width and transparency values to unravel the intricate interplay between these parameters and their impact on attack effectiveness.

Table 1: Results of ablation experiments

TS	W=0.1		W=0.3		W=0.5		W=0.7		W=0.9	
	ASR(%)	Query	ASR(%)	Query	ASR(%)	Query	ASR(%)	Query	ASR(%)	Query
0.1	40.07	342.86	51.47	279.15	56.62	254.54	55.88	252.63	55.88	238.96
0.2	74.63	180.78	82.35	119.05	86.03	95.29	88.24	82.90	87.13	84.38
0.3	93.75	82.39	96.69	41.27	97.79	27.61	98.53	23.08	98.90	15.62
0.4	96.69	46.17	99.63	15.72	100.00	8.55	100.00	4.25	100.00	4.15
0.5	98.53	35.68	99.26	13.68	100.00	4.72	100.00	2.93	100.00	2.25
0.6	98.53	29.50	100.00	10.78	100.00	3.50	100.00	2.38	100.00	1.64
0.7	98.53	25.11	99.63	10.04	100.00	3.02	100.00	1.88	100.00	1.36
0.8	98.53	23.85	99.26	9.44	100.00	2.79	100.00	1.62	100.00	1.28
0.9	98.16	25.06	99.63	6.81	100.00	2.15	100.00	1.67	100.00	1.23



Figure 4: Digital samples

Table 1 presents a comprehensive overview of the outcomes derived from the digital ablation experiments. A thorough scrutiny of these experimental findings allows the formulation of the following four pivotal conclusions: Firstly, it is evident that ADCP consistently demonstrates proficient digital attack outcomes across a spectrum of experimental configurations. Remarkably, even in the most challenging scenario (W=0.1, TS=0.1), ADCP still achieves a commendable attack success rate of 40.07%. Secondly, there exists a discernible correlation between the augmentation of both the width and transparency of the camera patch and the amplification of ADCP's attack success rate. This alignment with expectations is mostly upheld, except for the isolated case (W=0.1, TS=0.9). It is pertinent to mention that, in the experimental context, a higher value of the transparency parameter TS actually signifies a diminished transparency of the camera patch. Thirdly, a pinnacle attack performance is realized by ADCP when the camera patch's width is set to 0.5 and the transparency to 0.4. This signifies that further incremental adjustments in either width or transparency do not significantly elevate the attack success rate, as ADCP has already attained a 100% success rate in this configuration. Finally, an in-depth

examination of the outcomes delineated in Table 1 unveils that ADCP attains robust attack performance when the transparency TS reaches 0.3.

Based on the aforementioned insight, the configuration range for physical attacks has been delimited as follows: $0.1 \leq \mathrm{W} \leq 0.3, 0.3 \leq \mathrm{TS} \leq 0.5$. Furthermore, in other digital attack experiments, the width has consistently been set to 0.1 and the transparency to 0.5. Figure 4 graphically presents the adversarial samples engendered by the approach with a width of 0.1 and a transparency of 0.5. The top row showcases the detection outcomes for clean samples, while the subsequent row showcases the detection outcomes for adversarial samples. This visual confirmation substantiates the efficacy and adversarial potency of the methodology. These salient conclusions yield profound insights into the research, fostering an enhanced understanding of ADCP's attack characteristics, and furnishing pragmatic guidance for the configuration of physical attacks in real-world applications.

Physical test: To comprehensively assess the approach's efficacy in a physical context, the physical experiments are divided into two phases: indoor testing and outdoor testing.

Table 2: Results of indoor test experiments (ASR)

Distance	4.6m	5.2m	5.8m	6.4m	7.0m	7.6m	8.2m	8.8m
pl60+pl60	77.11%	88.03%	92.89%	83.07%	45.45%	55.96%	89.19%	83.00%
pl60+w57	98.02%	90.12%	92.13%	98.58%	98.53%	97.82%	99.50%	97.57%

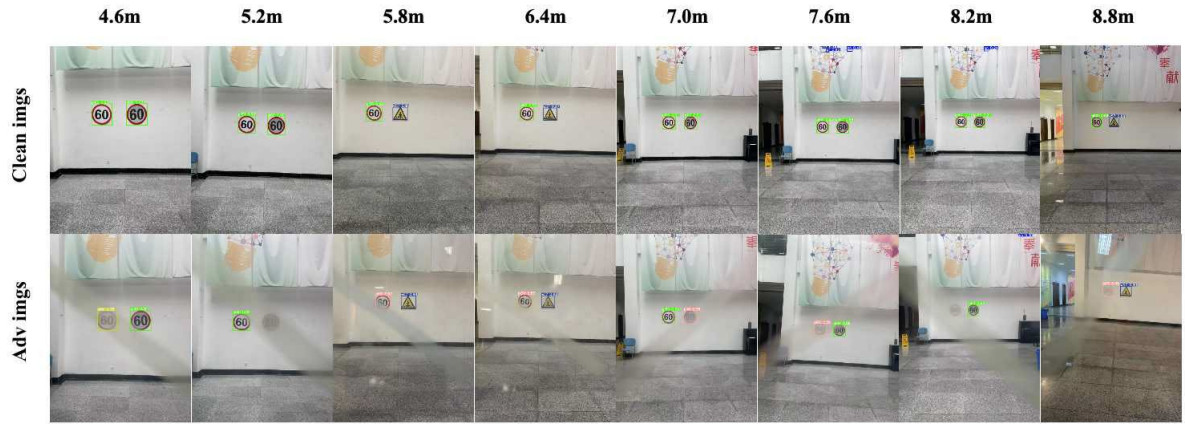


Figure 5: Indoor test

Indoor Testing Phase: To comprehensively assess the adversarial impact of the method across varied conditions, two distinct sets of experiments are conducted during the indoor testing phase. In the initial set, identical road signs are employed, while the subsequent set involved different road signs. These experiments are video-recorded, with the resultant footage being divided into frames to generate physical samples. Subsequently, the ASR is calculated based on these samples. The findings reveal that in the two sets of experiments, we accumulate 1630 and 4772 physical samples, respectively, achieving attack success rates of 78.16% and 96.31%. To gain a deeper understanding of ADCP's adversarial effect across different distances, a distance-based subdivision of the attacks is conducted, and statistical analysis is performed. The results of this analysis are presented in Table 2. These outcomes unequivocally underscore the efficacy and resilience of the method across the various tested distances within the indoor testing context. To present these results in a more intuitive manner, Figure 5 showcases the physical samples employed in indoor testing phase. This visual representation emphasizes that the method effectively executes physical attacks irrespective of the distance involved. During this testing stage, a camera patch width of W=0.1 and a transparency of TS=0.3 are opted for. Of notable importance is the fact that the presence of perturbations remains nearly imperceptible to the human eye without meticulous observation. This underscores the inherent imperceptibility of the attack method. Through this series of indoor testing experiments, confidence in the method's effectiveness within a controlled environment has been substantially fortified, laying a robust foundation for the research.

Table 3: Experimental results of outdoor testing(ASR)

Distance	4m	5m	6m	7m	8m	9m	10	11m	12m
0°	56.88%	100.00%	60.28%	61.26%	98.11%	100.00%	84.34%	87.50%	100.00%
30°	100.00%	100.00%	85.85%	87.16%	100.00%	72.90%	100.00%	100.00%	59.05%
45°	79.61%	63.03%	78.39%	73.58%	93.41%	100.00%	76.36%	87.96%	55.56%

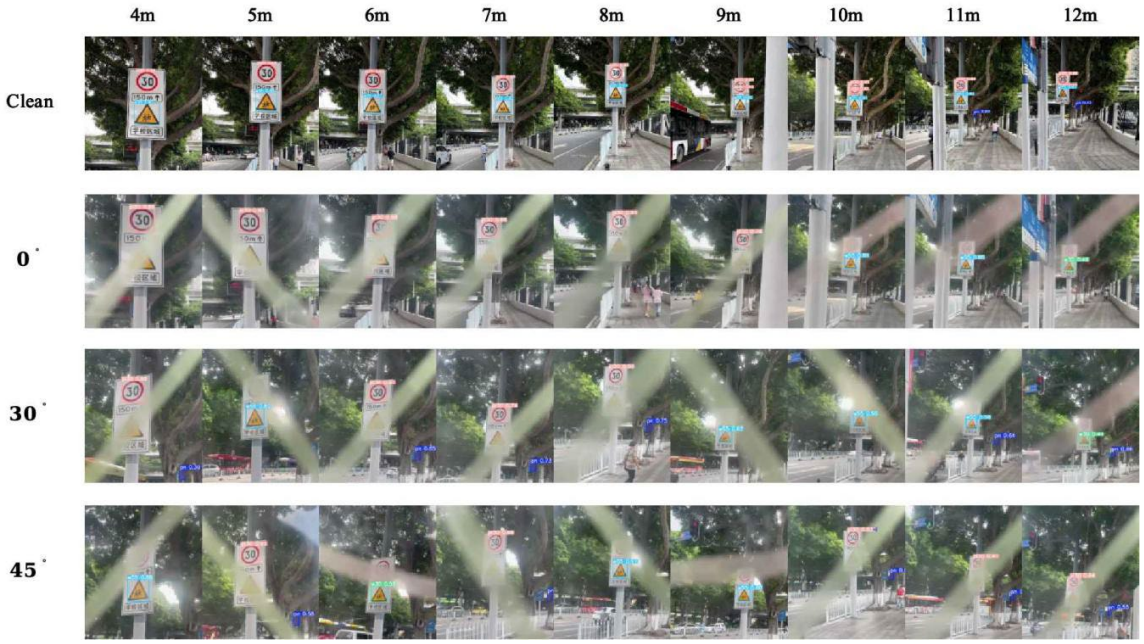


Figure 6: Outdoor test

Outdoor Testing Phase: In the outdoor testing phase, a transition is made to utilizing genuine road signs to subject the method to realistic attack scenarios. This shift allowed for a more comprehensive assessment of the method's robustness in real-world conditions. The objective is to evaluate the effectiveness of the method across various distances and angles, mirroring the unpredictability of genuine outdoor environments. During the outdoor testing, the method adhered to an attack configuration of $0.1 \leq \mathrm{W} \leq 0.3$ and $0.4 \leq \mathrm{TS} \leq 0.5$, thereby striking a balance between attack effectiveness and stealthiness. Impressively, outdoor testing yielded a total of 2624 genuine physical samples(pl30+w55), culminating in a remarkable attack success rate of 83.31%. This notable outcome underlines the practical viability and adversarial prowess of the approach within outdoor settings. The detailed experimental findings are elaborated in Table 3, which unequivocally illustrates that the method consistently manifests effective and resilient physical attack outcomes across diverse distances and angles. Remarkably, one third of the distance-angle combinations resulted in a 100% attack success rate. Even in the worst scenario (12m, 45°), a noteworthy attack success rate of 55.56% is achieved. Figure 6 visually captures the adversarial samples generated by ADCP during outdoor testing, further underscoring the method's capacity to execute impactful physical attacks across varying distances and angles. This comprehensive validation confirms the robustness and applicability of the method in complex outdoor scenarios. The series of outdoor experiments has yielded highly favorable outcomes, enhancing confidence in the practical applicability of the approach within real-world contexts.

Table 4: Comparison of experimental results between the approach and baselines

Method	Scenario	Digital attack		Physical attack	
		ASR	Query	Indoor ASR	Outdoor ASR
AdvCS(Li et al, 2019)	White-box	49.60%	ø	ø	73.26%
TTP(Zolfi et al,2021)	White-box	42.47%	ø	ø	42.27%
ADCP（Ours）	Black-box	93.75%	83.39	78.16/96.31%	88.31%

Table 4 presents a comprehensive comparison of experimental outcomes between ADCP method and the baseline approach. By scrutinizing these comparative findings, ADCP method consistently exhibits a more robust adversarial effect in both digital and physical settings when contrasted with the baseline method. This outcome underscores the inherent superiority of the approach. It is imperative to acknowledge that ADCP method operates in a black-box scenario, aligning closely with real-world practical applications. Moreover, in contrast to the AdvCS and TTP methodologies that employ multiple camera patches, which consequently yield larger experimental errors and inadequate stealthiness, ADCP approach's utilization of a solitary camera patch yields a higher attack effectiveness coupled with superior stealthiness. Drawing from the comprehensive analysis of Tables 1 and 4, it can be confidently assert that ADCP method bears a more formidable realistic threat compared to AdvCS and TTP methods. Its impressive experimental performance not only attests to its efficacy but also underscores its potential to pose significant challenges to computer vision systems within authentic environments. ADCP, thus, stands as a significant augmentation to the field of camera-based physical attacks, deserving attention for its multifaceted capabilities.

4.3 Ablation Study

In this section, a detailed exploration of another important influencing factor is conducted the color of the camera patch. The focus specifically encompasses the RGB channel of the color, encapsulating the *r, g,* and *b* values that span the range of (0, 255). Nevertheless, it is evidently impractical to execute ablation experiments for every conceivable color scenario, given the potential explosion in the number of experimental combinations. To strike a balance in the experimental design, a judicious trade-off was opted for. Ablation experiments with three distinct values were performed for each of the RGB channels: 0, 127, and 255. This approach resulted in 27 experimental configurations, carefully designed to thoroughly explore the role of color in the effectiveness of ADCP attacks. Figure 7 illustrates the outcomes of these color ablation experiments.

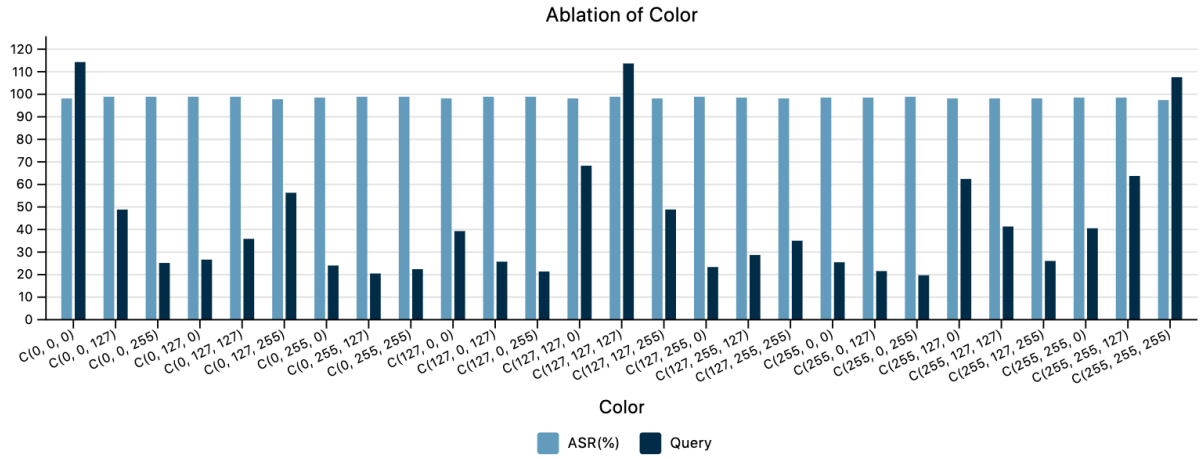


Figure 7: Ablation of C

The most important observation from these results is that color has minimal impact on the adversarial effectiveness of ADCP. In essence, camera patches sporting diverse colors seemingly exert relatively minor sway over the potency of the attack. This discovery tangibly bolsters the robustness and universality of the method. The method's ability to maintain a relatively steadfast attack effectiveness across different color variations reinforces its utility and adaptability, thus underscoring its reliability in contexts.

5. Conclusion

In this study, we introduce ADCP, a novel camera-based physical attack, optimized using the PSO algorithm. ADCP's key advantage over traditional methods lies in its ability to achieve successful attacks without modifying the target object, enhancing stealthiness and providing an advantage over existing patch-based and light-based attacks. Additionally, ADCP demonstrates superiority in operational simplicity and robustness, using only one camera patch to achieve reliable and resilient black-box attacks. The comprehensive experimental framework and results demonstrate the effectiveness and resilience of ADCP in real-world scenarios, evidencing its substantial adversarial impact in both digital and physical realms, as well as its high success rates in physical testing. The method's invisibility stems from perturbing the camera itself, enhancing stealth. The research highlights ADCP's security threats in the physical domain, with implications for enhancing comprehension and

application of physical attacks. With the increasing prevalence of physical adversarial attacks, the search for effective countermeasures continues to be a pivotal research focus.

Acknowledgements

This research work was partly Sponsored by Natural Science Foundation of Xinjiang Uygur Autonomous Region(No.2022D01B185).

References

Athalye et al.(2018) Synthesizing robust adversarial examples, in: International conference on machine learning, PMLR. pp. 28

Chen et al.(2019) Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18, Springer. pp. 52–68.

Goodfellow et al.(2014) Explaining and harnessing adversarial examples. arXiv:1412.6572 .

Guesmi et al.(2023) Advrain: Adversarial raindrops to attack camera-based smart vision systems. arXiv preprint arXiv:2303.01338.

Gnanasambandam, A., Sherman, A.M., Chan, S.H.(2021) Optical adversarial attack, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 92–101.

Hu et al.(2023) Adversarial laser spot: Robust and covert physical-world attack to dnns, in: Asian Conference on Machine Learning, PMLR. pp. 483–498.

Hu et al.(2022) Adversarial texture for fooling person detectors in the physical world, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13307–13316.

Huang et al.(2021) An improved shapeshifter method of generating adversarial examples for physical attacks on stop signs against faster r-cnns. Computers & Security 104, 102120.

Jocher(2020) Ultralytics yolov5. URL: https://github.com/ultralytics/ yolov5, doi:10.5281/zenodo.3908559.

Kennedy and Eberhart(1995) Particle swarm optimization, in: Proceedings of ICNN'95-international conference on neural networks, IEEE. pp. 1942–1948.

Li et al.(2019) Adversarial camera stickers: A physical camera-based attack on deep learning systems, in: International Conference on Machine Learning, PMLR. pp. 3896–3904.

Suryanto et al.(2022) Dta: Physical camouflage attacks using differentiable transformation network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15305–15314.

Szegedy et al.(2013) Intriguing properties of neural networks. arXiv:1312.6199 .

Tan et al.(2021) Legitimate adversarial patches: Evading human eyes and detection models in the physical world, in: Proceedings of the 29th ACM international conference on multimedia, pp. 5307–5315.

Thys et al.(2019) Fooling automated surveillance cameras: adversarial patches to attack person detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp. 0–0.

Wang et al.(2021) Dual attention suppression attack: Generate adversarial camouflage in physical world, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8565–8574.

Wu et al.(2020) Making an invisibility cloak: Real world adversarial attacks on object detectors, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer. pp. 1–17.

Zhong et al.(2022) Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15345–15354.

Zhu et al.(2016) Traffic-sign detection and classification in the wild, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2110–2118.

Zolfi et al.(2021) The translucent patch: A physical and universal attack on object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15232–15241.

A Case Study on a Multi-Country Money Laundering Scheme and a Proposed Automatic Detection System

Tsz-Fung Tony Tse[1,2], Xiao Tan[1,2], Siu Ming Yiu[2] and Hiu-Man Human Lam[3]
[1]Hong Kong Police Force, China
[2]Dept of Computer Science, University of Hong Kong, China
[3]Department of Justice, Hong Kong, China

tonytse@police.gov.hk
xtan@cs.hku.hk
smyiu@cs.hku.hk
humanlam@doj.gov.hk

Abstract: This paper presents a case study on the Franco-Israeli syndicates orchestrating their cross-continental money laundering schemes. These money laundering schemes have been operating for over two decades, ever since China's entry to the World Trade Organization in 2001. The paper reviews the operation of the money laundering schemes in detail and highlights the difficulties encountered by bankers and investigators in unearthing and investigating criminal activities within the banking systems. The paper then proposes an automatic anti-money laundering system, which is expected to address these difficulties. Preliminary experimental results show that the system successfully identifies the crux of these money laundering syndicates within a few days' time, something which usually takes years of investigation to track down the suspects using traditional methods, as well as its ability to initiate pre-warning procedures to the banks and law enforcement agencies once suspicious transaction clusters are found. The paper concludes with a discussion on the legal implications encompassing the evidence projected by this system.

Keywords: Cross-continental money laundering, Investigation difficulties, Legal challenges, Automatic AML detection

1. Introduction

1.1 Background

Taking advantage of the "One Country, Two Systems" principle, Hong Kong is crowned world's most open economy. The city operates under a sophisticated financial and legal infrastructure and benefits from various economic freedoms, including the free flow of capital, a free trade policy, a simple taxation system and low tax rates, and the advanced transportation and telecommunication systems, attracting global investors from all around the globe. In 2014, around a quarter of Mainland China's trades chose Hong Kong as the intermediary, be it in the form of offshore trade or re-exports (Hong Kong Monetary Authority, 2016).

Unlike Hong Kong, Mainland China has imposed a foreign exchange control which prevented money from moving in and out of the country freely. Despite the capital controls, China's economy and international trade volumes have been growing substantially and is now the world's second largest economy (Worlddata, 2022). There is a growing need for capital flow, be it for legal or illegal purposes. This encourages the use of underground banks and cash smuggling across the Chinese borders, with Hong Kong, a world-renowned free market economy, being the most viable option of all.

1.2 Introduction of Anti-Money Laundering

Anti-money laundering ("AML") is a major research topic in FinTech; money laundering is a main funding source for organized crimes. Today, money laundering has evolved into an algorithmic-based system with millions of offshore/onshore accounts, traversing countries and continents.

One appropriate definition of money laundering was introduced in Masciandaro 1993, which is structured by two key-characteristics (Donato Masciandaro. 1999): illegality including organized crime, drug trafficking and terrorist financing (general feature) and concealment to hide the illegal source of such revenues (specific feature).

There are three steps in money laundering: placement, layering and integration. Placement is placing the illegal funds in another form, which will enable the money launderer to undertake further layering and therefore disguise these amounts. Layering is washing the placed suspicious funds to produce a false source of fund, and thereby disguising the proceeds of crime, such that the original source and the current position of the funds are unclear (i.e. investing the placed funds into something legitimate). The integration is the final stage, which is to integrate the after-layering assets into a legitimate financial system, with the profits or

benefits earned via the layering process, and to transfer the liquidity assets, which can then be converted to cash easily, all done within the legitimate financial systems.

1.3 Our Contributions

In this paper, the tactics of the trade-based money laundering by the Franco-Israeli syndicates are introduced. Using international trades as its façade, it starts off with the purchase of goods or commodities in Eastern Europe, goods of which are then shipped to and sold in Hong Kong and China, the money, from the selling these goods or from the purchase of antiques or art collections, was then sent to Israel. With the infamous mastermind of such organized money laundering groups being a French-Israeli (Timesofisrael, 2016), such trade-based Anti-money laundering cases are hence called "France-Israeli syndicates". In this case study, we will (i) provide insights and observations as to the difficulties in detecting and investigation this type of money laundering cases; (ii) propose a more efficient automatic detection system which can identify similar suspicious money laundering cases efficiently, significantly shortening investigation by a years' time; and (iii) discuss the legal implications on and the admissibility of the evidence generated from the automatic system.

We hope that the results of this article will ultimately benefit the community in terms of the technological developments of AML tools, the criminal investigations carried out by law enforcement agencies, and the billing of legal frameworks or legislatures by the judiciaries.

2. Selected Money Laundering Cases

Before we provide the observations and insights on the Franco-Israeli model of money laundering (ML), three relevant cases will be outlined to illustrate their ML schemes.

2.1 Case A

In February 2009, an Israeli male – Mr. A came to Hong Kong, incorporated four companies and opened business bank accounts for these companies. Mr. A left Hong Kong shortly after the bank accounts were opened. Ten months later, in December 2009, one of the bank accounts was used to receive crime proceeds of HKD 1.9 million, which originated from a fraud case against a bank in France. The case was later reported to the Police by the bank.

Between June 2009 and January 2010, the police investigations revealed that a total of HKD 69.1 million was remitted from France, Germany and Portugal to the bank accounts set up by Mr. A, and was then further transferred to various locations including China, Cyprus, Israel and Hungary. When Mr. A was arrested, he stated that these bank accounts were opened for tax evasion purposes only. In the end, Mr. A was charged with money laundering.

2.2 Case B

In June 2009, an Israeli male – Mr. B came to Hong Kong, incorporated one company called Golden Longon Ltd. and opened a bank account at the "Hong Kong and Shanghai Banking Corporation" ("HSBC"). Mr. B left Hong Kong shortly after the bank account was opened. One month later, in January 2010, a French company was deceived in a telephone deception scam and transferred Euro 982,000 to Mr. B's account.

Bank record of the HSBC account showed that two transfers, totaling Euro 1.48M was made to a local Bank of China ('BOC") account under the name of Bisley Global BVI Co. in June 2009 and November 2009 respectively. Bisley Global BVI Co. was incorporated by another suspect, Mr. C in April 2009. The company had a total of three accounts held at BOC, Standard Chartered Bank ("SCB") and HSBC and Mr. C was the sole account signatory. Bisley Global BVI Co. was reported to be engaged in the trading of computers in Hong Kong and overseas.

Mr. C had also incorporated a local company – Tobo Group Ltd. in March 2009. The company had a total of two bank accounts at BOC and SCB and Mr. C was the sole account signatory. Tobo Group Ltd. was reported to be engaged in general trading in Hong Kong and overseas. Altogether, Mr. C had five business bank accounts, three under the name of Bisley Global BVI Co. and two under the name of Tobo Group Ltd.

Bank records showed that large funds were deposited into bank accounts operated by Mr. C, there were some inter-transactions among these bank accounts. And funds were mainly withdrawn either by way of remittances or were withdrawn in cash. Fund flow showed that most of the fund originated from France, Germany and Cyprus, and was then further transferred to Mainland China, Israel or back to Cyprus. Although funds were drained out eventually, the disposals were not all done immediately. In fact, some funds were held

in the accounts for weeks and months before they were transferred. Between May 2009 and April 2012, at least HKD 120M was laundered via these bank accounts.

Investigation into the bank accounts, under the name of Nice Alpha Ltd (incorporated by Mr. C), found that Euro 1.52M & USD 0.3M were transferred from a local BOC account under the name of a local company - Company Z - between April 2010 and September 2010. The company had another bank account at SCB. A local male – Mr. L (Mr. C's friend) was the account signatory of both accounts. At least HKD 31.05 million was laundered. The fund flow and the operation period of the two bank accounts were similar to the bank accounts operated by Mr. C, i.e., majority of the fund was originated from France, Germany and Cyprus, etc. and was further transferred to Mainland China, Israel or back to Cyprus. Some of the funds were first transferred to local money service operators before they were transferred to Mainland China. Investigation with the recipients in Mainland China of this case revealed that those transfers were supposed to be monies sent from the senders' relatives, who were working in Europe. In the end, Mr. B was included in the Interpol Red Notice.; Mr. C and Mr. L were charged with money laundering.

2.3 Case C

Mr. D is an Australia and Israel national, but regularly resided in Israel. Between January 2010 and April 2012, by instruction of his supervisor in Israel, Mr. D came to Hong Kong on 21 occasions, incorporated 9 companies and opened the business bank accounts for these companies. Mr. D knew that these bank accounts were used to receive crime proceeds (or "black money").

During the two years, Mr. D used the "black money" to purchase 342 gold bars from a gold trader and stored them in two village houses. He also collected three 5-carat diamonds from a jewelry trader on three occasions, for which he was not required to make any payments. On a final note, Mr. D also received HKD 2.68M in his personal bank accounts at HSBC and Citibank.

In May 2012, three other Israeli males – Mr. E, Mr. F and Mr. G were employed by an unknown male (not Mr. D) to assist Mr. D in removing the gold bars and diamonds from the two village houses. Some gold bars were sold back to the gold trader who sold the gold bars to Mr. D in the first place. The sale proceeds, with a total of HKD 73 million, were paid to the company bank accounts controlled by Mr. D.

Mr. D – Mr. G were all charged with money laundering and convicted for the offences charged (South China Morning Post, 2014).

3. Observations and Insights

According to the Hong Kong Money Laundering and Terrorist Financing Risk Assessment Report in 2022, over half of the convicted Money Laundering cases between 2016 and 2020 were identified as third-party Money Laundering ("ML") cases, and bank accounts are still one of the most common tools exploited by money launderers for both domestic and foreign predicate offences via cash deposits, cash withdrawals or wire transfers (Financial Services and the Treasury Bureau, 2023). Prolonged offence period, use of MSO (Money Service Operator), use of multiple layering and structuring, and use of various shell companies and bank accounts all added difficulties in identifying and investigating this money laundering model.

4. A Proposed Automatic AML Detection System

Traditionally, the investigation of such cases is long and resource consuming. Indeed, it took the Hong Kong Police more than one year to deduce the money laundering processes of the above three cases. The traditional way of doing it is to enter all entries in to an excel table including the date of transaction, the amount involved, the parties involved, the beneficiary owner of the bank account, etc. and to identify out the common linkage and money laundering pattern from it. It is not just time consuming, there is also likelihood for the investigation unit to miss out a linage by overlooking the data. However, with the support of Machine Learning and Network Analysis, and the open-source tool (e.g.: Gephy), the process of deduction was shortened substantially. The proposed automatic anti-money laundering detection system can identify suspicious transactions efficiently and accurately and can detect the masterminds, using the self-learn capabilities of artificial intelligence (AI).

4.1 The State-Of-The Art AML using Artificial Intelligence

As early as 2007, the study to automate the Anti-Money Laundering Process is proposed for finance institutions (Han, Jingguang, et al., 2020). Machine learning methodologies, e.g. classification and clustering,

are widely applied to AML systems, in the detection of suspicious transactions. In 2007, Gao and Xu proposed the Herbert A. Simon's Model of Decision-Making, which was used to calculate the likelihood of the transactions being that of high-risk AML (Gao, Shijia, and Dongming Xu., 2009). In 2012, Stefan Axelsson *et. al.* introduced a classification method for money laundering detection in a data set consisting of synthetic financial transactions, thereby targeting anomalies inside a data set of mobile money financial transactions, by using the classification techniques to categorize transactions as either suspicious or non-suspicious (Lopez-Rojas, Edgar Alonso, and Stefan Axelsson, 2016). In the paper published by Nida S. Khan *et. al.* in 2013, a Bayesian network (BN)-based approach is proposed to calculate the account holder's transaction behavior score, based on the transaction history. An alert will be generated if a significant difference is detected in the account holder's transactional patterns and behaviors (Khan, Nida S., et al., 2013). However, for the aforesaid supervised methods, the imbalanced training dataset between normal transactions and anomalous transactions is the key defect, in terms of the performance in detecting suspicious transactions.

Except for the machine learning and deep learning methodologies of transactions identification and classification, network analysis is another type of methodology used in analyzing the money laundering process. Common network analysis systems contain variables, e.g., the degree of centrality, authoritativeness, and closeness centrality. These indices measure the closeness and betweenness of the relationship of the connections and the importance of each node in the networks. In 2015, Drezewski *et. al.* introduced a network analysis of bank statements and the National Court Register data to deduce the analysis of social networks for money laundering cases. In 2017, Colladon and Remondi (2017a, b) built several networks to work collectively, including transactions, the economic sector, geographical area, and tacit link networks to prevent money laundering. They used the 19-months data of a factoring company that mainly operates in Italy. They found that network metrics were extremely useful in fraud-risk assessment (Xiao, Guohui, et al., 2019).

In our Automatic Anti-money Laundering System design, both the anomalous transaction detection technology and the link analysis are adopted and combined to construct the main architecture of the system. Except for the suspects detection and monitoring agents, which is based on the unsupervised algorithms and the link analysis, the knowledge graph technology works as the reference library and event analysis to refresh and supplement the features on accounts and account holders' level. The details would be described in the sections below.

4.2 The Methodology

In this section, we propose an Automatic Anti-Money Laundering Detection System ("AAMLDS") which combines the unsupervised algorithm with the link analysis on the financial institutions' transactions, accounts and account holders' data. The data is kept updated using the knowledge graph to identify the missing information or potential errors, as well as to discover the new suspect personalities and their *modus operandi*. The architecture design of the system is briefed as below.

The original data sources include the transaction, accounts and account holders' data from financial institutions, as well as the reference data from external authorities and the AML-related news reports. The original data are loaded to the AAMLDS with an API agent. Based on the different purposes of the transactions, there are two components in AAMLDS. The first one is the Anomalous Transactions Detection Agent, which adopts the incremental principal component analysis ("IPCA"), or density-based spatial clustering of applications with noise (" DBSCAN") for detecting the anomalous changes with imbalanced distributed and unlabeled transaction data. This helps identify suspicious accounts or account holders to generate the labelled data from the original unlabeled raw data. Then, based on the detected suspicious accounts and account holders, the Network Analysis Agent builds up the link analysis and social network analysis ("SNA") to deduce the complete money laundering fund flow by creating the visual representations of transactions to track the movement of the funds, thereby identifying the source and the destinations of the funds, illustrating the relationships between accounts, and predicting the trends and patterns of money laundering (Dre˙zewski, Rafał, Jan Sepielak, et al., 2015). In this paper, we will take Amazon Web Services ("AWS") as the cloud platform example.

This system adopts both security orchestration, automation, and response (SOAR) and elastic security info and event management (SIEM) to build up the security operations platform (Kinyua, Johnson, and Lawrence Awuah., 2021). SIEM investigates and alerts the security risk by analyzing the logs and events. Such traditional security approach is enhanced by SOAR, an automatic, software-defined security with AI/ML involved incidents responses (Sridharan, Anish, and V. Kanchana., 2022). Such security operation platform is capable to provide enhanced security protection for the internal and external data transmission, storage and analysis in the cloud.

This system combines the Anti-money Laundering databases from financial institutions with the alert reports and analysis to the regulatory and the law enforcement agencies. Those reports are refreshed regularly to provide the complete and updated funds flow-in and flow-out processes between various suspects and suspicious accounts. The pros of the system include:

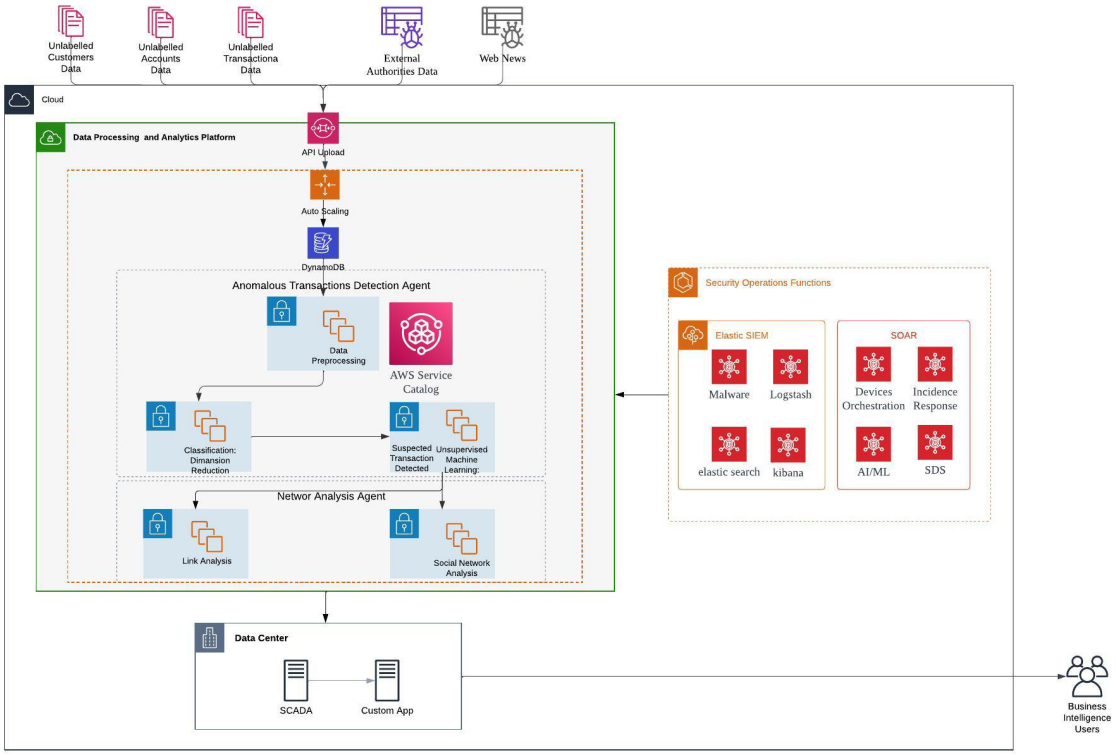


Figure 1: Architecture Design of Automatic Anti-Money Laundering Detection System

4.2.1 Transaction data collection and pre-processing of banking records

The automatic AAMLD system starts from analyzing the suspicious transaction records reported or collected from financial institutions, including banks, insurance companies, trustees, and other investment entities. Generally speaking, 12-36 months of transactions in the accounts are required for training purposes. Basic filtering and pre-processing are deployed to standardize the data formats and to validate the integrity of the data quality from different systems.

4.2.2 Unsupervised algorithms of features extraction

With or without the single label of classifications of the transaction data, unsupervised algorithms are deployed to detect the anomalous transactions from the trained normal transactions. As one type of clustering algorithms, the IPCA based DBSCAN is applied to extract the features of anomalous transactions without the requirement of the predefined labels of the trained dataset.

There are quite a few challenges for solving the Anomalous Transactions Detection problem, and those challenges are mainly related to the limitations of the data collections, which include:

- Imbalanced training dataset.

For most applications, anomaly events with a long time-series data are rare. Generally speaking, data files rarely contain a large volume of normal data and actual anomalous transactions because of the complex scenarios outlined in our case examples. IPCA based DBSCAN, as an unsupervised arbitrary center-based clustering approach, utilizes the density reach ability and the density connection ability to find the optimized center and the radius of each non-overlap clusters when training the unlabeled input data, and detect the anomalous points based on the outliers of any existing clusters of the training.

Because it disregards the normalcy of the dataset input, unbalance will not impact the performance of the algorithm in nature.

- Stability of high performance.

To maintain the high accuracy of detection and prediction, its performance is another issue in real-life application. For example, the Convolutional Generative Adversarial Network and Recurrent Convolutional Autoencoder are capable of achieving high accuracy after sufficient training. However, the stability of the overall predictive performance for each anomality detection is still a concern (Fensel, Dieter, et al. 2020).

Unlike the convolutional neural network etc. deep learning algorithms, IPCA based DBSCAN, as one of the density-based clustering, applies a local cluster criterion. Because of the different methods for regulatory and optimization process, after conducting dimensions reduction using IPCA, DSBCAN adopts the linear optimized calculation of the density and radius and is thus able to achieve the high accuracy once it finds all the clusters after training.

- Hard-coded architecture.

Because different transactions have their own unique signatures, therefore the localized models are required for each different transaction's types or *modus operandi*. DBSCAN in this paper adopts the parameterized architecture instead of hardcoded when defining the radius from the center point to the boundary point ("EPS") and the minimum count of points to define a cluster ("MinPts"). Such self-adapted architecture is able to reduce the dimensionality to the calculated framing for the algorithm to proceed. For different transactions' types, the algorithm will change the temporal of the input data automatically.

Such design, on some level, helps to build auto-localized models to improve the self-adaptability of different money laundering patterns.

4.2.3 Network analysis and data visualization

Based on the results of anomalous transactions detection, such transactions and accounts holders would be filtered out as the entities to set up the network analysis and deduce the money laundering patterns.

The indices of network analysis include: the degree of centrality, which indicates the number of direct connections of one node/vertex in the network; authoritativeness, which is the degree to which one node points to another node via important hubs, and betweenness centrality, which is the measure of the extent to which an actor has control over information flowing between others. During the network analysis procedure, the roles are assigned to nodes in the network, the roles' connections are measured, the entities' mutual proximity are attempted to be determined, and this information to the external role information (e.g., bank statements and the National Court Register) assigned to nodes are compared.

In this article, link analysis and social network analysis are adopted to deduce the process of the money laundering crossing Europe, Hong Kong and China, and identify the roles of the suspicious accounts and account holders in the process. With the auto-construction of the relations between suspicious accounts/account holders, the Network Analysis Agent is able to remove the noise, deduce the money laundering pattern and identify the relationships between suspects. This will be discussed in detail in the results of the experiments.

4.3 The Experiments

4.3.1 The setup

The training data is collected from Kaggle.com (website: https://www.kaggle.com/datasets/ealaxi/paysim1), which only keeps normal transactions. The suspicious transactions are the real data collected from the three cases mentioned in the article, including the eight accounts of Bisley Global, Nice Alpha and Tobo Group in different local banks with different currencies in Hong Kong.

The data classification for anomaly detection is based on the transactions data, including the transaction types, the transaction amount, the balance between the transaction, the balance after the transaction etc. The transaction types include Cash-in, Cash-out, Payment, Debit and Transfer.

In the experiment, the training dataset only includes the normal data downloaded from Kaggle.com, and we randomly pick up 25,000 transactions in the total 65,499 sample normal in 50 times consecutively. The test dataset is from the HK Police, in which we only selected the five attributes to keep consistent with the training dataset. The test dataset includes 200 suspicious transaction records monitored by the banks, which were randomly selected from the total 395 suspicious transaction records and 200 normal transactions data,

totaling at 400. The prediction focuses on the turning point from consecutive 25,000 normal transactions data to 400 mix-up transactions data. Both the train and the test datasets are scaled and normalized.

4.3.2 Performance comparison and analysis

AUC, PAUC, F measure, Accuracy Score, MSE, and Spearman Rank Correlation Coefficient etc. would be set up as the evaluation matrix for performance measures. From the experiment result of 50 random test cases shown in Table 1, it can be observed that when detecting the suspicious transactions of Bisley Global's accounts, the average accuracy is about 0.82, the stability indicators, including Jaccard Score and Spearman Rank Correlation Coefficient are 0.64 and 0.67 respectively. And the detection performance of the suspicious transactions of Bisley Global, which is 26.23% higher than that of Tobo Group. From the result of the experiments, the transactions of Bisley Global are more suspicious for money laundering than those of Tobo Group.

Bisley Global is sentenced for money laundering.

Table 1: Anomaly detection between two suspects

Measurements	Bisley Global	Tobo Group
Number of Test Cases	50	50
AUC	0.82	0.65
PAUC	0.77	0.61
Accuracy	0.82	0.65
MSE	0.18	0.35
Spearman Ranking Correlation Coefficient	0.67	0.37
Jaccard Score	0.64	0.33

4.3.3 Social network analysis

The output generated by the Anomalous Transaction Detection is loaded for the Network Analysis. Based on the classified anomalous transactions and the refreshed data correction and complement from knowledge graph engine, the link analysis and the social network analysis are deployed to set up the relationships between the suspicious accounts and the account holders, and to visualize the money laundering process and pattern.

The social network analysis focuses on the accounts and accounts' holders (customers). And the data classification includes transactions related, accounts related and customer related. Transactions related attributes include Sum-Counts of Transaction, Sumamt and DRCR. Accounts related attributes include target account's ID, target currency, target bank, target region/country, source account's ID, source currency, source bank and source region/country. And Customers related attributes include target name of account holder(s) and source name of account holder(s).

In this case, there are 24 source bank accounts and 86 target bank accounts which are also related to the findings of the Hong Kong Police.

Based on the link analysis of the suspicious accounts shown in Figure 2, several key points could be observed:

- "Bisley Global BVI Co." account of EUR in Bank of China in Hong Kong is the central of the transactions between all accounts.
- "Nice Alpha" is the second centralized entity in the network. The "Nice Alpha" of EUR in Standard Chartered Bank in Hong Kong only receives/sends money with the corporate accounts in Europe and in Hong Kong. The "Nice Alpha" of EUR in Bank of China in Hong Kong transacts with both the individual accounts in China and the corporate accounts in Hong Kong, China and Europe.
- "Tobo Group" of EUR in Standard Chartered Bank in Hong Kong is recognized as another centralized one in the constructed linked network. Similar to the "Nice Alpha" of EUR in Standard Chartered Bank in Hong Kong, this account also transacts only with the corporate accounts in Hong Kong and Europe.

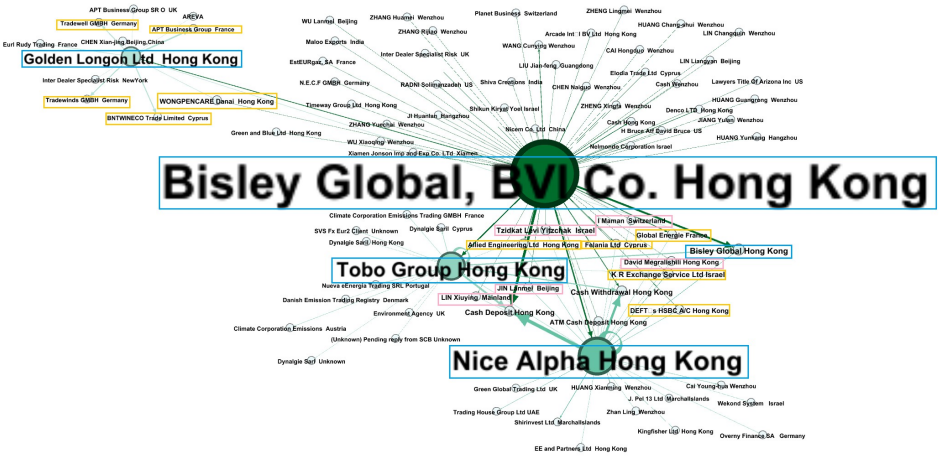


Figure 2: Link Analysis of Suspicious Accounts' Holders

Figure 2 is the link analysis for the suspicious account holders, and the transactions between different accounts are merged if those accounts belong to the same account holder. It shows that there is a cone in the central of the network with the nodes of Bisley Global Hong Kong, Nice Alpha Hong Kong and Tobo Group Hong Kong. In the cone, the direction of the fund is from Bisley Global to Alpha and then to Tobo Group. Except for the transferring out from Nice Alpha to Tobo, there are two other exits: Cash Withdraw and Cash Deposit.

"Golden Longon Limited" of EUR in Bank of China in Hong Kong is observed as another centralized node in the network, but they layered small percentage of the laundered money in the network and only transacted with business accounts in France, Cyprus, or German in Europe.

Bisley and Nice Alpha are charged with money laundering.

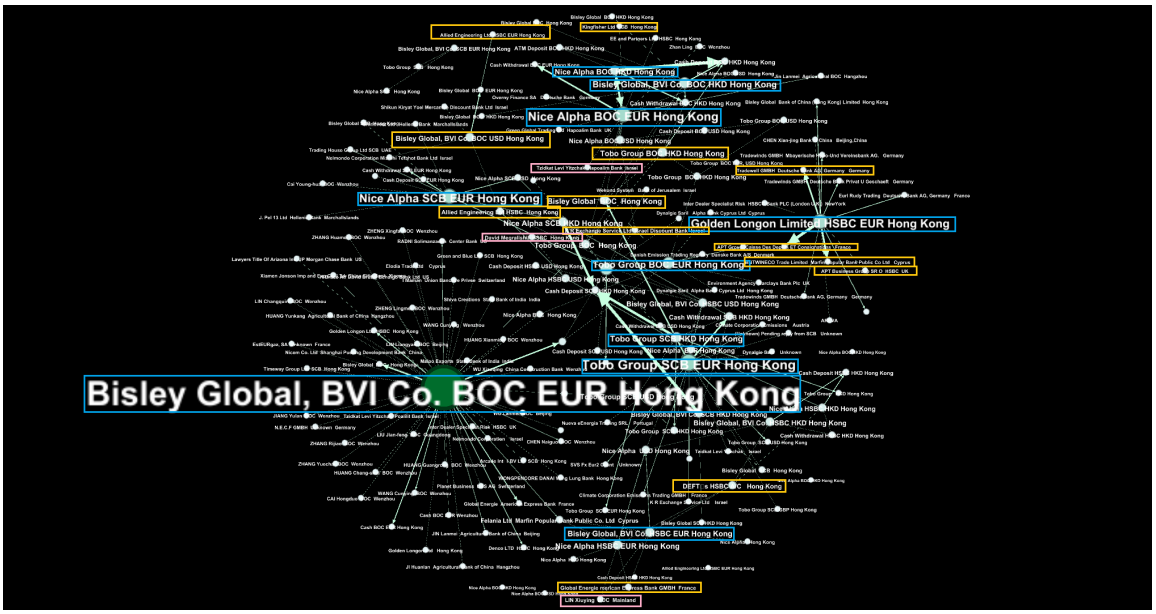


Figure 3: Social Network Analysis of Suspicious Accounts

From the relationship analysis of customers, we extend such analysis to suspicious accounts. Figure 3 shows the hierarchy structure of the accounts, a three tier accounts as follows: -

1. The first-tier accounts are the centralized nodes in the social network. These accounts all belong to the four suspicious account holders: Bisley Global, Nice Alpha, Tobo Group, and Golden Longon. From the narrows of each node of the first tier, it can be observed that these nodes are the main initiators and the main intermediary of the fund flows, because of their high degree centrality and betweenness centrality of other normal nodes. These first-tier accounts are highlighted with the blue frames in Figure 3.

2. Second tier accounts are the accounts which transact with the first-tier accounts more frequently than other accounts. The edges between the first and the second-tier accounts are bolder than those between the first tier and other normal accounts, and most of them are linked to multiple first level accounts. The main observations of second level accounts include:

- Most of the corporation accounts of the second level are from Cyprus, France, Israel, Germany and Hong Kong, e.g.: Felania Ltd Marfin Popular Bank Public Co. Ltd in Cyprus, BNTWINECO Trade Limited Marfin Popular Bank Public Co Ltd in Cyprus, Global Energie American Express Bank GMBH in France, and APT Group Caisse Des Depots ET Consignations in France and UK, etc.
- Three individual accounts work as the connections between Bisley Global and Nice Alpha, and these accounts are from Hong Kong, Israel and Mainland.

3. Finally, the third-tier accounts are the ultimate recipient of the fund, or at least very remote from the first tier account.

5. Conclusion and the Discussion in Future

Automated analysis has yet been widely applied or presented as evidence in criminal proceedings and thus is rarely discussed by the Hong Kong Courts. This section highlights some of the legal implications concerning the admissibility of automated analysis in money laundering criminal cases. In order for the automated analysis to be admissible, one of the criterions must be the genuineness of the underlying raw data such as the inter-account transfers, over-the-counter cash transactions, etc. These raw data should be admissible evidence, supported by a banker's affirmation, under section 20 of the Evidence Ordinance, Cap 8, Laws of Hong Kong (Hong Kong e-Legislation, 2018).

For the purpose of section 20 of Evidence Ordinance, "*any document or record used in the ordinary business of a bank*" includes any report of meeting between the account holder and his banker and credit card sales slip.

Secondly, the software used to generate the automated analysis must be proven to be reliable. The developer should have pre-established models and criteria that are specific and effective, making it possible to identify hallmarks of money laundering in financial activities. The models and criteria should be subject to regular re-examination to ensure that they are reliable and up to date (Westlawasia, 2021).

Since automated analysis should have no or minimal margin of error in processing the raw data, alternatively, the result from the automated analysis should be subject to independent examination by 'non-automated measures' before being tendered for use as evidence in criminal proceedings.

Before employing the software in performing automated analysis, law enforcement agencies may also consider carrying out tests on the software to determine whether it is accurate. Such a test would then provide evidence that the software would then be regarded as reliable in producing automated analysis (Westlawasia, 2012).

In the experiments of anomalous transaction detection, it is observed that the performance of unsupervised anomalous transactions detection achieves 0.82 and is capable to identify the suspicious object from other objects in the real case. However, the volume and complexity of the raw data for the experiments can be improved to mimic the real anti-money laundering environment.

Besides, based on the anomalous transactions' detection engine and the network analysis engine, we expect to set up the knowledge graphs to deduce and enrich the completed money laundering process. Such processes could be mapped to and validated by the admissible evidence, and to be accepted by the court and law enforcement. With knowledge graph engine, the efficiency and accuracy of both anomaly detection and network analysis can be improved continuously. We expect to study on this domain and add more features of Automatic Anti-Money Laundering Detection System in future.

References

Donato Masciandaro. (1999): "Money Laundering: the Economics of Regulation", European Journal of Law and Economics, 7:225–240.

Dre˙zewski, Rafał, Jan Sepielak, et al. (2015): "The application of social network analysis algorithms in a system supporting money laundering detection." Information Sciences 295: 18-32.

Fensel, Dieter, et al. (2020): "Introduction: what is a knowledge graph?" Knowledge Graphs. Springer, Cham, 1-10.

Financial Services and the Treasury Bureau (2023): Hong Kong Money Laundering and Terrorist Financing Risk Assessment Report 2022, Money Laundering Report_2022_EN.pdf (fstb.gov.hk)(accessed 08 Feb 2023).

Gao, Shijia, and Dongming Xu. (2009): “Conceptual modeling and development of an intelligent agent-assisted decision support system for anti-money laundering.” Expert Systems with Applications 36.2: 1493-1504.

Han, Jingguang, et al. (2020): “Artificial intelligence for anti-money laundering: a review and extension.” Digital Finance 2.3: 211-239.

Hong Kong e-Legislation. (2018): Section 20 of the Evidence Ordinance, Hong Kong, Cap. 8 Evidence Ordinance (elegislation.gov.hk)(accessed 09 Feb 2023).

Hong Kong Monetary Authority (2016): Hong Kong, The Global Offshore Renminbi Business Hub, Hong Kong The Global Offshore Renminbi Business Hub (hkma.gov.hk), (accessed 09 Feb 2023).

Khan, Nida S., et al. (2013): “A Bayesian approach for suspicious financial activity reporting.” International Journal of Computers and Applications 35.4: 181-187

Kinyua, Johnson, and Lawrence Awuah. (2021): "AI/ML in Security Orchestration, Automation and Response: Future Research Directions." Intelligent Automation & Soft Computing 28.2.

Lopez-Rojas, Edgar Alonso, and Stefan Axelsson (2016): ”A review of computer simulation for fraud detection research in financial datasets.”2016 Future technologies conference (FTC). IEEE.

South China Morning Post (2014): Israeli soldier convicted for helping money-laundering syndicate, https://www.scmp.com/news/hong-kong/article/1508586/israeli-soldier-convicted-helping-money-laundering-syndicate

Sridharan, Anish, and V. Kanchana. (2022): "SIEM integration with SOAR." 2022 International Conference on Futuristic Technologies (INCOFT). IEEE.

Timesofisrael (2016): How a French-Israeli grifter became a money-laundering pioneer in China, https://www.timesofisrael.com/how-a-french-israeli-grifter-became-a-money-laundering-pioneer-in-china/

Westlawasia (2021): La Quadrature du Net & Ors v Premier Ministre & Ors (2021) 1 WLR 4457. https://launch.westlawasia.com/document/I9C19D40024BD11ECA3D5D6E91F41AA26?srguid=i0ad832f20000018c2b32c34c6e617a61&fromSearch=true&offset=1

Westlawasia (2012): Bevan v Western Australia (2012) WASCA 153. https://launch.westlawasia.com/document/I549a0ff0d7c811e2a455b12f5e340c12?startChunk=1&endChunk=1&fromSearch=true&offset=1

Worlddata (2022): The world's largest economies, https://www.worlddata.info/largest-economies.php.

Xiao, Guohui, et al. (2019): “Virtual knowledge graphs: An overview of systems and use cases.” Data Intelligence 1.3: 201-223.

Cyber Operations in Peace and War: A Framework for Persistent Engagement

Brett van Niekerk
Durban University of Technology, South Africa
Security Insitute for Governance and Leadership in Africa, Stellenbosch University, South Africa

brettv@dut.ac.za

Abstract: During 1990s, the concept of information warfare (IW) and information operations (including cyber operations, psychological operations, and electronic warfare) could be conducted with varying intensity across all stages of peace and conflict. At that time, many of the concepts related to cyber operations were still hypothetical. Subsequently, conflicts and competition between states have demonstrated the capabilities and limitations of cyber operations. Research emerging in 2022 by multiple authors demonstrate the limitations and usage of offensive cyber operations and maintaining a sustainable military cyber capability, as well as proposing alternative models for conflict in cyberspace. Alongside this, there has been increased attention on the impact of ICTs on international security and the responsible behaviour of nation-states in cyberspace. There is still ambiguity and different perspectives on the application of international law in cyberspace. This uncertainty disrupts the original models of IW and warfare, which assumed clear distinctions amongst the conflict stages. Both the discourse of ICTs in international security and recent conflict necessitate a reconsideration of the decades-old view of IW in times of peace, war and the grey zone in between. This paper proposes a framework for the roles of cyber operations across the stages of conflict based on contemporary perspectives on the utility of cyber operations as well as practical examples. In rethinking the IW model, a multidisciplinary view is required, considering the technical, legal, social and international security perspectives.

Keywords: Cyber warfare, Influence operations, Information operations, Information warfare, Persistent engagement

1. Introduction

During the advent of information warfare (IW) and information operations in the 1990s, operations (including cyber operations and psychological operations) could be conducted to varying degrees through all stages of peace and conflict. While the scope of many IW operations was proposed for each of the conflict stages, at the time many of the concepts related to cyber operations were still hypothetical. Subsequently, the capabilities and limitations of cyber operations have been demonstrated (for example, in the Russo-Ukraine conflict). Research by Moore (2022), Smeets (2022), and Fischerkeller, Goldman and Harknett (2022) demonstrate the limitations and usage of offensive cyber operations and maintaining a sustainable military cyber capability.

In addition, there has been an increase in discussion on the impact of ICTs on international security, the relevance of international law to cyberspace, as well as responsible behaviour of nation-states in cyberspace. While there is agreement that international law applies to cyberspace and a set of norms have been established, there is still ambiguity on how international law is to be translated from the physical context to the virtual environment; for example, perspectives differ on what the threshold of an armed attack or an act of war in cyberspace is. This uncertainty disrupts the original models of IW, which assumed clear distinctions amongst the conflict stages. Furthermore, cyber diplomacy arising from the above-mentioned forums provides another mechanism of state power related to cyber operations, but also creates complexities.

Both the discourse of ICTs in international security and the Russo-Ukraine conflict necessitate a reconsideration of the decades-old view of IW in times of peace, war and competition. This paper proposes a framework for the roles of cyber operations across the stages of conflict based on contemporary perspectives on the utility of cyber operations, as well as lessons emerging from Ukraine and other practical examples. In rethinking the IW model, a multidisciplinary view is required, considering the technical, legal, social and international security perspectives.

Section 2 presents a background to facets of information warfare, the stages of conflict, and international cybersecurity. Cyber and disinformation operations are discussed in Section 3 with a view of demonstrating how they evolve over various stages of conflict. Section 4 provides the discussion and proposes an updated model for cyber operations across the stages of conflict; this is followed by the conclusion in Section 5.

2. Literature Review

This section provides the background to IW (Section 2.1), the stages of conflict (Section 2.2), and cyber diplomacy and cyber operations in international security (Section 2.3).

2.1 Information Warfare

Information warfare is a concept from the 1990s that incorporated a number of information activities aimed at affecting decision making and military operations (Waltz, 1998; Jones, Kovacich and Luzwick, 2002); the concept could be expanded into peacetime and was known as information operations. However, the usage of these terms has changed and tend to focus on the disinformation aspects, or occasionally purely the cyber component. Originally, information warfare included command and control warfare, military deception, economic information warfare, cyber operations, psychological operations, network-centric warfare, intelligence, and electronic warfare (Waltz, 1998; Jones, Kovacich and Luzwick, 2002). In addition, psychological operations have evolved with a strong cyber component in the form of influence operations due to the prevalence of social media. This evolution of mass disinformation and influence through the use of social media, supported by cyber operations, can be seen as a tool for diplomatic (or other) pressure on citizens and governments. Key to this is election interference, which has evolved due to these practices, and illustrated by Rid (2020). While Europe exhibits a number of examples of foreign influence operations, approximately 30 recorded disinformation operations in Africa are considered to originate externally, with half attributed to a single country (Africa Centre for Strategic Studies, 2022).

2.2 Stages of Conflict

Brahm (2003) lists the stages of conflict as (illustrated in Figure 1): latent conflict, conflict emergence, escalation, stalemate, de-escalation, settlement, and peacebuilding. However, Brahm (2003) notes that some stages may be repeated, and the transition between stages may not be smooth. For example, many conflicts may attempt to reach a settlement, only for an incident to occur and result in further escalation. The broader situation in the Middle East can be considered as one that oscillates amongst various stages; similarly, the situation in Ukraine de-escalated for a period, before returning to a full conflict scenario.

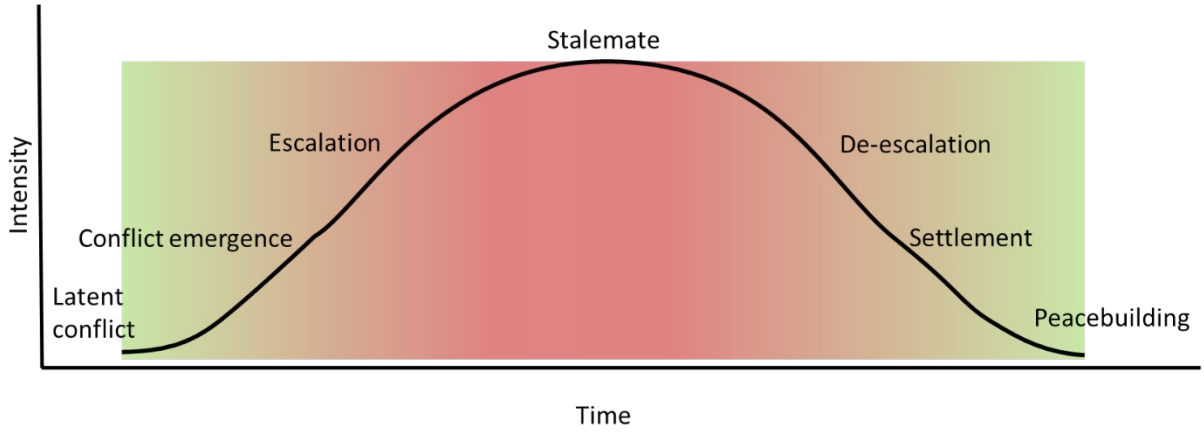


Figure 1: Stages of conflict, adapted from (Brahm, 2003)

Jones, Kovacich and Luzwick (2002) and Waltz (1998) present a similar six stage model for conflict, and indicate the prevalence of various IW activities for the stages, as is illustrated in Figure 2. The six stages are: considered as intelligence gathering, diplomatic pressure, economic pressure, military posturing, combat, and reconstruction.

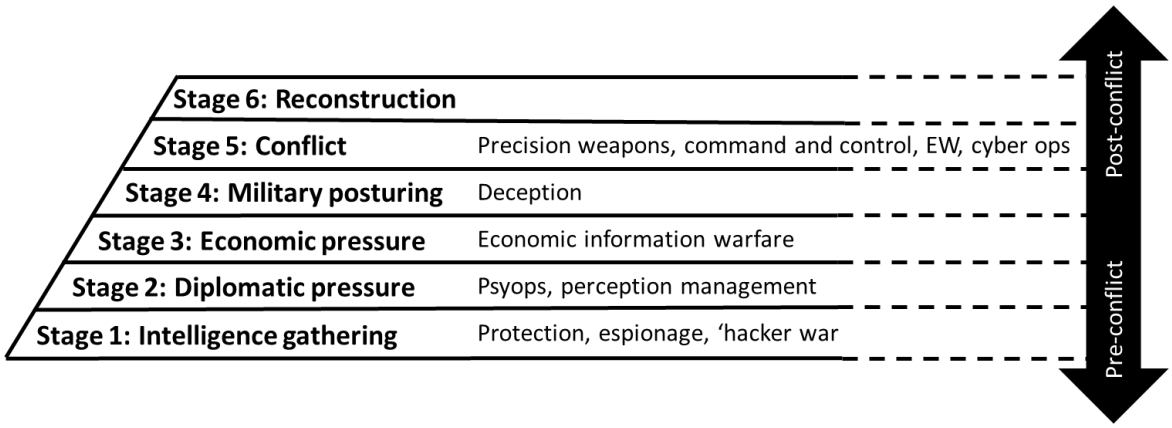


Figure 2: Stages of IW, adapted from Jones, Kovacich and Luzwick (2022) and Waltz (1998)

These models, however, are dated, and modern conflict appears to exhibit 'persistent' activity in the cyber domain. Therefore, the purpose of this paper is to propose a new alignment of cyber activities across conflict stages based on the concept of persistent engagement. For the purposes of this paper, the stages of conflict from both models need to be considered in conjunction with one another. The two models can be roughly aligned as shown in Table 1. It should be noted that diplomatic and economic pressure can be used as a tool for de-escalation, where pressure on the parties can result in them in seeking a settlement.

Table 1: Mapping of the two conflict models

Brahm (2003)	**Jones, Kovacich and Luzwick (2022) and Waltz (1998)**
Latent conflict	Intelligence gathering
Emergence	Diplomatic and economic pressure
Escalation	Military posturing
(Hurting) Stalemate	Conflict
De-Escalation	Diplomatic and economic pressure
Settlement/Resolution	Diplomatic and economic pressure
Post-Conflict Peacebuilding and Reconciliation	Reconstruction

2.3 Cyber Diplomacy and Cyber Operations in International Security

In attempt to address the potential impacts of cyber operations in an international security setting, a series of Group of Governmental Experts (GGEs) were established in 2004 under the auspices of the United Nations; a total of six GGEs were held, ending in 2021 (Schmitt, 2021). Of note, the 2015 GGE proposed non-voluntary norms to guide nations on responsible behaviour in cyberspace (UNGA, 2015). In addition to the GGE processes, Open Ended Working Groups (OEWGs) have been established as a more inclusive approach (Digital Watch, 2023). A number of initiatives have arisen in parallel to the UN processes, such as the Paris Call for Trust and Security in Cyberspace (2018) which proposed nine principles, and the Global Commission on the Stability of Cyberspace (2019), which proposes norms to protect the public core of the Internet. These initiatives are termed cyber diplomacy, which can be defined as "the means by which nations, groups, or individuals conduct their affairs in cyberspace, in ways to safeguard their interests and promote their political, economic, cultural or scientific relations, while maintaining peaceful relationships" (EU Cyber Diplomacy Toolbox, 2023). Cyber diplomacy helps guide state cyber operations and the implications for international security.

Healey (2021a) categorises four main frameworks for cyber operations in international relations research:

- Dyadic – cyber operations between specific organisations
- Strategic – cyber operations between nations through campaigns
- Asset – specific to technology (e.g. software and network)
- Systemic – system of systems related to the Internet

Research into the escalatory nature of cyber operations was conducted by Lonergan and Lonergan (2023) and Buchanan (2017), where Buchanan frames the escalatory nature within the context of a security dilemma in

cyberspace. Maness, Valeriano, Hedgecock, Macias, and Jensen (2023) consider the dyadic model of cyber incidents. Chesney and Smeets (2023) edited a book discussing cyber operations as primarily an intelligence competition. The concept of cyber persistence theory (persistent engagement) were discussed in Fischerkeller and Harknett (2019) and Fischerkeller, Goldman, and Harknett (2022).

Smeets (2022) discusses the challenges of establishing and maintaining an effective military cyber capability (as opposed to entering the cyber arena). Moore (2022) considers cases of a number of nations and their offensive cyber operations, and he notes that Iran recognises that achieving a top-tier status in the cyber domain is unlikely, therefore the focus is on being good enough to achieve objectives and signal that they are able to do so. However, Smeets (2022) indicates that a country with emerging capability that is determined to use it no matter what, may prove to be more dangerous than a state with developed capability but is restrained in its use.

Two posture statements released in 2023, one by the UK National Cyber Force (2023) and the other by General Nakasone (2023) of the US Cyber Command, provide insights into the operations conducted by the governmental cyber units. In both instances, the postures align to that of persistent engagement. However, both these nations are considered as having established capability and responsible legal guidance on their operations; other nations may have different capabilities and postures, and this could ultimately alter the dynamics thereby affecting the engagement in cyberspace (Healey, 2021a; Smeets, 2022).

Cyber persistence theory is based on the fact that the interconnectedness of nations through the Internet allows for competition in cyberspace, where cyber operations below the threshold of an armed attack can be conducted with limited risk of escalation. In some instances, the concept of forward defence may also be relevant, where a nation can disrupt on adversary's capability to conduct operations – in essence this is to defeat a cyber or influence operation as close to the source as possible. (Harknett, Fischerkeller & Goldman, 2023). Buchanan (2017) also considers the forward defence, but places this in the context of a security dilemma which results in escalation dynamics. A key component proposed by Fischerkeller and Harknett (2019) is that there is agreed competition. The current discussion in the cyber diplomacy circles can form part of this agreed competition even if it is implicit: national perspectives on international law in cyberspace are provided, which indicates the 'line' which could be considered escalatory.

3. Cyber Operations and Influence Operations Across the Stages of Conflict

This section provides perspectives on where and how cyber operations been employed during both periods of conflict and during times of peace.

3.1 The Early Days: Estonia and Georgia

In 2007 Estonia was subjected to distributed denial-of-service (DDoS) attacks in protest of a proposed relocation of a Soviet World War 2 memorial. Media, government websites and email, and banks were affected (Landler & Markoff, 2007; Rolski, 2007). This incident was the initial demonstration of the potential for cyber operations to be used for political and economic pressure.

In 2008, Russian forces intervened in the Georgia-South Ossetia conflict. Prior to this, DDoS attacks targeted the Georgia government and communications, resulting in a communications blackout. In addition, website defacements were used for propaganda and psychological operations (Coleman, 2008; Hart, 2008). This incident was a preview of how cyber operations could support military operations by creating a communications blackout. This was later refined for use in Ukraine, as describe in the next section.

3.2 Ukraine

Ukraine provides a particularly useful illustration of cyber operations across various stages of conflict. In 2014 popular protests ousted a pro-Russian government, which was followed shortly afterwards by the Russian occupation of Crimea, and an internal conflict in the East with pro-Russian separatists. In early 2022, Russian forces entered Ukraine. Cyber operations have played a part of this conflict; the CyberPeace Institute (2023) has recorded 2929 cyber operations by 122 unique threat actors associated with the Russo-Ukraine conflict between January 2021 and October 2023.

The incidents up to the end of 2022 are described in more detail by van Niekerk (2023), and a summary is provided below. In 2013 to 2014, cyber operations and online disinformation campaigns were conducted to support website defacements; DDoS attacks targeted regional organisations, but were prominent in attempting to disrupt Ukrainian communications prior to the Russian military incursion into Crimea, the Crimean referendum and the 2014 Ukrainian elections (also targeted by a disinformation campaign). Between the two

periods of physical operations, notable cyber operations included the disruption of the Ukrainian power grid in 2015 and 2016 by Russian advanced persistent threat (APT) groups, the NotPetya malware which is the costliest malware incident, a Ukrainian water treatment plant, and continued disinformation campaigns (van Niekerk, 2023). In addition, between 2014 and 2016 Ukrainian mobile phones were infected with spyware allowing Russian actors to track Ukrainian artillery units (van der Waag-Cowling, van Niekerk, & Ramluckan, 2022).

Immediately prior to the Russian incursion, an increase in disinformation campaigns was seen. The satellite provider Viasat was targeted to impact Ukrainian communications (similar to in Georgia and the Crimean annexation). During the early stages, notable cyber incidents included attempts by hacktivists to disrupt trains carrying Russian troops, and operations targeting Ukrainian power stations. A number of cyber operations targeting Russia were more symbolic, such as 'hack and leak' operations that released information about a nuclear power plant. What was unique during this period was the foreign assistance in the form of 'hunt forward' operations by US cyber command, and foreign corporations providing technical support to Ukraine. In addition, the Ukrainian president openly called for a volunteer cyber army to aid Ukraine (van Niekerk, 2023).

Subsequent reports indicate that there was a successful attempt on the Ukrainian power grid in October 2022, which coincided with missile strikes (Nelson, 2023), with ongoing attempts against the Ukrainian energy, government, and telecommunication sectors (CyberPeace Institute, 2023). The targeting appears to be more towards civilian infrastructure, to erode the national resilience and will, rather than to achieve specific military objectives.

The Ukrainian example demonstrates the used of influence operations to psychologically prepare the battlefield, and cyber operations to disrupt communications in the early stages of military action. Coordination between cyber operations and kinetic attacks appears to be evolving, however limited incidents appear to have the opportunity to directly impact on the battlefield, but rather targeting broader national and civilian resilience.

3.3 Iran

In 2010, the malware known as Stuxnet was discovered, and the prime target was determined to be the Iranian nuclear enrichment facility at Natanz. This cyber operation was significant in that it was the first recorded to affect industrial control systems and cause physical damage (Zetter, 2014); in addition, it demonstrated the strategic use of cyber operations n an attempt to maintain the status quo in terms of the balance of nuclear power.

Moore's (2022) opinion is that the Stuxnet incident made Iranian authorities realise the potential for cyber operations to further their asymmetric efforts at regional power projection; while they are not at the stage of a top-tier cyber power, their capabilities have evolved so that they are 'good enough' to achieve their objectives (Moore, 2022; Smeets, 2022). Notable cyber operations attributed to Iran (or proxy groups) include a wiper known as Shamoon targeting Saudi Aramco (Moore, 2022); the 2012 DDoS attacks against US banks and financial institutions either in protest to the movie *Innocence of Muslims* (Harris, 2014) or in response to US sanctions (Council on Foreign Relations, 2023); the targeting of a US dam (Brandom, 2016); the 2013 targeting of an unclassified US Marine Corps network and the 2014 targeting of US casino chains, Shamoon 2 and then in 2018 Shamoon 3 which targeted an Italian firm contracted to Saudi Aramco (Moore, 2022); and the targeting of the UK parliament in 2017 (The Telegraph, 2017). In July 2022, cyber operations targeted Albanian networks; these were attributed to Iran, resulting in Albania severing diplomatic ties with Iran in September of that year (Reuters, 2022). Iran was also accused, alongside Russia and China, in conducting influence operations and attempting to influence the 2020 US presidential elections; in 2017 Twitter reportedly closed 7000 fake accounts associated with Iranian influence operations (Timberg & Romm, 2019). Shample (2023) describes eight Iranian-linked APTs, all engaged with cyber-espionage against various regions and sectors.

Beyond the Stuxnet operation, the US alleged conducted cyber operations against Iranian air defence in response to the downing of a US drone in 2019 (BBC, 2019) and then again in response to an alleged Iranian attack on a Saudi Aramco oil facility (Ali & Stewart, 2019). In 2020, Iran allegedly attempted to target Israeli water infrastructure, to which Israel responded by targeting and disrupting the main Iranian port (Warrick & Nakashima, 2020; Times of Israel, 2020). In 2021, a hacktivist group targeted an Iranian steel plant, and then the petrol stations in 2022 (Bob, 2022).

The cyber operations both targeting and attributed to Iran demonstrate a range of objectives, from economic and diplomatic pressure, espionage, sabotage, to the possible disruption of air defence equipment. While none were conducted during the state of an official war, these incidents demonstrate cyber-operations across a range of conflict stages, with possible applications to a kinetic war. The cyber operations also demonstrate concepts

of proportional responses, where a cyber operation was conducted when a possible kinetic option was available, implying persistent engagement over escalation dynamics.

3.4 Election Interference

In 2016, foreign actors targeted the US presidential elections through a combination of disinformation and cyber operations, including a 'hack and leak' operation related to the Democratic National Committee emails. While these efforts were probably not directly successful; they did cast concern over election validity (amongst other scandals) and susceptibility (Office of the Director of National Intelligence, 2017).

As mentioned above, the 2014 Ukrainian elections were targeted by a cyber operation where the results were attempted to by altered, and this was supported by disinformation operations. However, the alteration was detected and unsuccessful, but the disinformation pre-empted the results and broadcast their altered voting counts (Clayton 2014).

Foreign election interference was experienced in at least five European countries from 2017 to 2018, including cyber operations and disinformation campaigns. Prior to the elections, European countries began with preparatory measures to defend the integrity of their election processes from interference (Brattberg & Maurer, 2018). Proposed measures to mitigate the interference operations include: international collaboration, public statements as deterrence, and legal measures (which can be considered related to cyber diplomacy); awareness, sensitisation, and training of the population, with responsible media engagement and empowerment; implement vulnerability assessments and incident response/contingency measures (Brattberg & Maurer, 2018).

In the run-up to the 2020 US presidential elections, four foreign states were attributed in running disinformation and/or cyber operations against the elections: Russia, China, Iran, North Korea (Timberg & Romm, 2019). Prior to election day in 2020, US Cyber Command collaborated with Microsoft to disrupt the TrickBot botnet to mitigate the threat of ransomware affection the elections (Vavra, 2020). In 2022, the US agencies proactively acted to disrupt foreign actors that allegedly intended to interfere with the US midterm elections (Nakasone, 2023). This is an additional possibility to mitigate election interference besides those measures proposed by Brattberg and Maurer. While the European measures were more aligned with traditional defensive mechanisms and diplomacy, the 2022 US approach is more aligned with persistent engagement.

3.5 Criminal Takedowns and Counter Terrorism

As is stated in the UK's Nation Cyber Force statement, they "may also use a combination of technical and information operations against hostile actors in a mutually supportive way, for example, to sow distrust in groups such as criminal gangs or terrorist cells" (NCF, 2023:13) and "Countering threats from terrorists, criminals and states using the internet to operate across borders in order to do harm in the UK and elsewhere" (NCF, 2023:20).

The statement by Nakasone (2023) also indicates that the US Cyber Command can "enable efforts" (Nakasone, 2023) by other government agencies and partners to disrupt online activities by cybercrime and extremist groups. The TrickBot example (Vavra, 2020) mentioned above is an example of such activities. However, Healey (2021b) suggests limitations on when an intelligence and military organisation such as Cyber Command should be used to disrupt criminal activity.

An interesting example is the 2019 Israeli response to persistent cyber operations by a sub-state actor, where they targeted the building that was identified as being used to conduct the operations with an airstrike (Borghard & Schneider, 2019). This was the first time a kinetic military response was invoked by cyber operations.

4. Cyber Espionage

Cyber espionage is probably the most common and persistent state-back from of cyber operations, which have been used against adversaries and allies alike. Harris (2014), Segal (2016) and Kaplan (2017) provide detailed histories of the evolution of cyber operations, and the prevalence of cyber espionage during conflict and peace is clear.

The next section provides a discussion on cyber operations during peace and war, and proposes an updated model for cyber operations across the various stages of conflict.

5. Discussion and Proposed Updated Model

From the examples described in Section 3, it can be seen there is a mixture of traditional defence strategies and persistent engagement strategies when considering cyber operations. Cyber diplomacy is an ongoing process through various forums. However, both the UK and US statements indicate that cyber operations can be employed to support foreign policy (NCF, 2023, Nakasone, 2023), implying that cyber operations themselves are a form of cyber diplomacy.

The Ukrainian example illustrates an escalation of influence operations and strategic cyber operations targeting civilian infrastructure. Some cases of strategic cyber operations were present between the two conflicts; however, these were limited and may have been experimental. Cyber espionage was present throughout the two periods of conflict and the intervening years. Foreign assistance in the form of forward defence proved to be a new evolution. By contrast, the Iranian example illustrates more posturing and targeting of civilian infrastructure as a form of economic and diplomatic pressure. The use of cyber operations in response to kinetic operations, and potentially impacting weapons systems have been reported but in limited cases. In general, proactive forward defence operations have been employed to disrupt foreign influence operations, possible election interference, cybercrime infrastructure and online terrorist activity. However, there should be careful consideration of when military or intelligence cyber capability is deployed against non-state actors, as discussed by Healey (2021b). Cyber espionage is ever present throughout the stages of conflict.

From the cyber operations considered above, 'conflict' in cyberspace does not necessarily follow distinct stages, but rather is a continuum with varying levels of intensity and types of operations. While cyber operations will peak in preparation for conflict or during actual conflict, there does appear to be an escalation on some operations during periods of diplomatic and economic competition. This may be to support foreign policy and diplomacy (i.e. signalling), of for possible experimentation to test cyber capabilities. Overall, while there is an increase in intensity during conflict, the majority of cyber operations do not appear to be specifically beneficial to kinetic battlefield operations, even though there have been coordinated cyber and kinetic operations. This view is supported in the analysis by van der Waag-Cowling, van Niekerk and Ramluckan (2022).

Cyber diplomacy is a continuous process, as is evidenced by the number of initiatives; during peace the intensity will be low to medium, but will increase along with growing competition and tensions. This is as a result of the increase in cyber operations alongside the increase in tension, necessitating an increase in the cyber diplomacy. Influence operations are also likely to be ever-present, with the intensity rising during times of competition (including elections as influence operations may overlap with election interference). There is a case where the intensity of influence operations may peak drastically during military posturing in order to stoke tensions or psychologically prepare populations for kinetic military operations. Cyber espionage is another ever-present form of cyber operations, and is of consistently medium to high intensity due to its utility in maintaining competition in economic, military, diplomatic, and research arenas, as well as in other crisis situations. Forward defence will only be present with low intensity during latent conflict, in defending against routine cyber espionage and he occasional cyber crime or cyber terrorist threat. This will peak drastically during military posturing to mitigate any preparatory offensive cyber operations, and may also form an aspect of foreign assistance, as was seen in Ukraine in 2022. Strategic operations will most likely only emerge during diplomatic or economic pressure, and will escalate as tensions increase, but sharply drop during de-escalation. Table 2 presents a proposed model for the varying intensities of various cyber operations across the traditional stages of conflict. The shaded areas for each of the five cyber operation types indicate the intensity, with one to five blocks available to represent a rating of one to five (maximum).

Table 2: Alignment of key themes to documents and proposed norms

Stage	Cyber diplomacy	Influence operations	Cyber espionage	Forward defence	Strategic operations
Latent conflict: Intelligence gathering					
Emergence: Diplomatic & economic pressure					
Escalation: Military posturing					

Stage	Cyber diplomacy	Influence operations	Cyber espionage	Forward defence	Strategic operations
Conflict & Stalemate					
De-Escalation: Diplomatic & economic pressure					
Settlement/Resolution: Diplomatic / economic pressure					
Post-Conflict Peacebuilding, Reconciliation & Reconstruction					

A challenge in this analysis is that there has not been a complete settlement or resolution phase after a kinetic conflict. A limitation of this perspective is that the capabilities of the nations are not explicitly considered. The examples described are all related to nations with developed capability. Conflict between nations with developing or limited capability may exhibit different uses of cyber operations.

6. Conclusion

The growing prevalence of cyber operations globally, and particularly amongst a set of nations, has necessitated a reconsideration of information warfare activities during various stages of conflict. An international relations model that has been proposed is one of persistent engagement. From the analysis of various examples, cyber operations can be seen to be conducted as a continuum across all stages of peace and conflict, with the intensity of various types of operations varying and generally peaking during (or immediately prior to) kinetic conflict. This supports the persistent engagement model for nations with an established cyber capability. A limitation of the study is that there is limited information on post-conflict peacebuilding, and on conflict between nations with developing cyber capabilities. Future work can expand on this study once a broader range of examples become available.

Acknowledgement

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Number 150720).

References

Africa Center for Strategic Studies. (2022) *Mapping Disinformation in Africa*, 26 April, [online], accessed 18 January 2023, https://africacenter.org/spotlight/mapping-disinformation-in-africa/

Ali, I., and Stewart, P. (2019). Exclusive: U.S. carried out secret cyber strike on Iran in wake of Saudi oil attack: officials, Reuters, 16 October, [online], accessed 18 December 2023, https://www.reuters.com/article/us-usa-iran-military-cyber-exclusive/exclusive-u-s-carried-out-secret-cyber-strike-on-iran-in-wake-of-saudi-oil-attack-officials-idUSKBN1WV0EK/

BBC. (2019). US 'launched cyber-attack on Iran weapons systems', 23 June, [online], accessed 18 December 2023, https://www.bbc.com/news/world-us-canada-48735097

Bob, Y.J. (2022). Iran's steel industry halted by cyberattack, Jerusalem Post, 28 June 2022, [online], accessed 18 December 2023, https://www.jpost.com/middle-east/iran-news/article-710522

Borghard, E.D., and Schneider, J. (2019). Israel responded to a Hamas cyberattack with an airstrike. That's not such a big deal, *The Washington Post*, 9 May, [online], accessed 18 December 2023, https://www.washingtonpost.com/politics/2019/05/09/israel-responded-hamas-cyberattack-with-an-airstrike-thats-big-deal/

Brahm, E. (2003) Conflict Stages, *Beyond Intractability*, September, [online], accessed 25 October 2023, https://www.beyondintractability.org/essay/conflict-stages

Brandom, R. (2016). How a DDoS campaign became an act of cyberwar, *The Verge*, 25 March, [online], accessed 18 December 2023, https://www.theverge.com/2016/3/24/11301876/ddos-iran-banks-dam-prosecution-indictment

Brattberg, E., and Maurer, T. (2018). Russian election interference: Europe's counter to fake news and cyber attacks, Carnegie Endowment for International Peace, 23 May, [online], accessed 18 December 2023, https://carnegieendowment.org/2018/05/23/russian-election-interference-europe-s-counter-to-fake-news-and-cyber-attacks-pub-76435

Buchanan, B. (2017) *The Cyber Security Dilemma: Hacking, Trust and Fear between Nations*. Oxford University Press: Oxford.

Chesney, R., and Smeets M. (2023). *Deter, Disrupt, or Deceive: Assessing Cyber Conflict as an Intelligence Contest*. Georgetown University Press: Washington, D.C.

Clayton, M. (2014). Ukraine election narrowly avoided 'wanton destruction' from hackers, *Christian Science Monitor*, 17 June, [online], accessed 8 November 2022, https://www.csmonitor.com/World/Passcode/2014/0617/Ukraine-election-narrowly-avoidedwanton-destruction-from-hackers

Coleman, K. (2008). Cyberwar 2.0 - Russia vs Georgia, DefenseTech.org, 13 August, [online], accessed 18 December 2023, https://www.military.com/defensetech/2008/08/13/cyber-war-2-0-russia-v-georgia

Council on Foreign Relations. (2023). Denial of service attacks against U.S. banks in 2012–2013, [online], accessed 18 December 2023, https://www.cfr.org/cyber-operations/denial-service-attacks-against-us-banks-2012-2013

CyberPeace Institute. (2023). Cyber threats, [online], accessed 18 December 2023, https://cyberconflicts.cyberpeaceinstitute.org/threats

Digital Watch (2023). UN OEWG and GGE, Geneva Internet Platform, [online], accessed 12 January 2023, https://dig.watch/processes/un-gge

EU Cyber Diplomacy Toolbox. (2023). What is Cyber Diplomacy? Cyber Risk GmbH, [online], accessed 11 December 2023, https://www.cyber-diplomacy-toolbox.com/Cyber_Diplomacy.html

Fischerkeller, M.P., and Harknett, R.J. (2019). Persistent Engagement, Agreed Competition, and Cyberspace Interaction Dynamics and Escalation, The Cyber Defense Review, Special Edition, 267-287.

Fischerkeller, M.P., Goldman, E.O., and Harknett, R.J. (2022). *Cyber Persistence Theory: Redefining National Security in Cyberspace*. Oxford University Press: Oxford.

Global Commission on the Stability of Cyberspace. (2019). *Advancing Cyber Stability: Final Report*, November. Available at: https://hcss.nl/global-commission-on-the-stability-of-cyberspace-final-report/

Harknett, R.J., Fischerkeller, M.P., and Goldman, E.O. (2023). U.K. National Cyber Force, Responsible Cyber Power, and Cyber Persistence Theory, LawFare Blog, 5 April [online], accessed 14 December 2023 from https://www.lawfaremedia.org/article/uk-national-cyber-force-responsible-cyber-power-and-cyber-persistence-theory

Harris, S. (2014). *@War: The Rise of the Military-Internet Complex*, Houghton Mifflin Harcourt Publishing: New York

Hart, K. (2008). Longtime Battle Lines are Recast in Russia and Georgia's Cyberwar, *The Washington Post*, August 14, p. D01. Available at https://www.washingtonpost.com/wp-dyn/content/article/2008/08/13/AR2008081303623_pf.html

Healey, J. (2021a). The Offense-Defence Balance in Cyberspace. Offensive Cyber Working Group, 20 April, [Youtube], accessed 2 December 2023 from https://www.youtube.com/live/XggIUBCwtTA?si=f2wWpMgDgphxpiHS

Healey, J. (2021b). When Should U.S. Cyber Command Take Down Criminal Botnets?, *Lawfare*, 26 April, [online], accessed 18 December 2023, https://www.lawfaremedia.org/article/when-should-us-cyber-command-take-down-criminal-botnets

Jones, A., Kovacich, G. L., & Luzwick, P. G. (2002). *Global Information Warfare: How Businesses, Governments, and Others Achieve Objectives and Attain Competitive Advantages*. Auerbach Publications: Boca Raton, London & New York.

Kaplan, F. (2017). *Dark Territory: The Secret History of Cyber War*, Simon and Schuster: New York.

Landler, M., and Markoff, J. (2007). Digital Fears Emerge After Data Siege in Estonia. The New York Times Online, 29 May, [online], accessed 14 April 2010, http://www.nytimes.com/2007/05/29/technology/29estonia.html?_r=1

Lonergan, E.D., and Lonergan, S.W. (2023). *Escalation Dynamics in Cyberspace*. Oxford University Press: Oxford.

Maness, R.C., Valeriano, B., Hedgecock, K., Macias, J., and Jensen B. (2023). Expanding the Dyadic Cyber Incident and Campaign Dataset (DCID): Cyber Conflict from 2000 to 2020, *Cyber Defense Review* 8(2), 65-89.

Moore, D. (2022). *Offensive Cyber Operations: Understanding Intangible Warfare*. Hurst Publishers: London.

Nakasone, P.M. (2023). 2023 Posture Statement of General Paul M. Nakasone, US Cyber Command, 7 March, [online], accessed 12 April 2023 from https://www.cybercom.mil/Media/News/Article/3320195/2023-posture-statement-of-general-paul-m-nakasone/

National Cyber Force [NCF]. (2023). The National Cyber Force: Responsible Cyber Power in Practice, Crown Copyright. Available at: https://www.gov.uk/government/publications/responsible-cyber-power-in-practice

Nelson, N. (2023). Sandworm Cyberattackers Down Ukrainian Power Grid During Missile Strikes, Dark Reading, 9 November, [online], accessed 7 December 2023, https://www.darkreading.com/ics-ot-security/sandworm-cyberattackers-ukrainian-power-grid-missile-strikes

Office of the Director of National Intelligence. (2017). Assessing Russian Activities and Intentions in Recent US Elections, Intelligence Community Assessment ICA_2017_01. Available at: https://www.dni.gov/files/documents/ICA_2017_01.pdf

Paris Call for Trust and Security in Cyberspace. (2018) The 9 Principles, 12 November, [online], accessed 5 January 2023, https://pariscall.international/en/principles

Reuters. (2022). Albania cuts Iran ties over cyberattack, U.S. vows further action, 7 September 2022, accessed 18 December 2023, https://www.reuters.com/world/albania-cuts-iran-ties-orders-diplomats-go-after-cyber-attack-pm-says-2022-09-07/

Rid, T. (2020) *Active Measures: The Secret History of Disinformation and Political Warfare*, New York: Farrar, Straus and Giroux.

Rolski, T. (2007). Estonia: Ground Zero for World's First Cyber War, *ABC News*, 17 May, [online], accessed 23 September 2009, http://abcnews.go.com/print?id=3184122

Schmitt, M. (2021). The Sixth United Nations GGE and International Law in Cyberspace, *Just Security*, 10 June, [online], accessed 27 January 2024, https://www.justsecurity.org/76864/the-sixth-united-nations-gge-and-international-law-in-cyberspace/

Segal, A. (2016). *The Hacked World Order: How Nations Fight, Trade, Manoeuvre, and Manipulate in the Digital Age*, Public Affairs: New York.

Shample, S. (2023). Iranian APTs: An overview, *The Middle East Institute*, 10 February, [online], accessed 18 December 2023, https://www.mei.edu/publications/iranian-apts-overview

Smeets, M. (2022). *No Shortcuts: Why States Struggle to Develop a Military Cyber-Force*. Hurst Publishers: London.

The Telegraph. (2017). Iran blamed for cyberattack on Parliament that hit dozens of MPs, including Theresa May, 14 October, [online], accessed 18 December 2023, https://www.telegraph.co.uk/news/2017/10/13/iran-responsible-cyberattack-british-parliament/

Timberg, C., and Romm, T. (2019). It's not just the Russians anymore as Iranians and others turn up disinformation efforts ahead of 2020 vote, The Washington Post, 25 July, [online], accessed 18 December 2023, https://www.washingtonpost.com/technology/2019/07/25/its-not-just-russians-anymore-iranians-others-turn-up-disinformation-efforts-ahead-vote/

Times of Israel. (2020). Israel behind cyberattack that caused 'total disarray' at Iran port – report, 19 May, [online], accessed 18 December 2023, https://www.timesofisrael.com/israel-said-behind-cyberattack-that-caused-total-disarray-at-iran-port-report/

United Nations General Assembly. (2015) Report of the Group of Governmental Experts on Developments in the Field of Information and Telecommunications in the Context of International Security, 17th Session, UN Doc A/70/174, 22 July.

Van der Waag-Cowling, N., van Niekerk, B., and Ramluckan, T. (2022). The Use and Potential of Cyber Weapons in Contemporary and Future Conflict, *Future Warfare and Technology: Issues and Strategies,* ed. RP Rajagopalan*, GP-ORF Series,* 97-106. Available at: https://www.orfonline.org/research/future-warfare-and-technologies/

Van Niekerk, B. (2023). The Evolution of Information Warfare in Ukraine: 2014 to 2022, *The Journal of Information Warfare* 22(1), 10-31.

Vavra, S. (2020). Cyber Command, Microsoft take action against TrickBot botnet before Election Day, CyberScoop, 12 October, [online], accessed 18 December 2023, https://cyberscoop.com/trickbot-takedown-cyber-command-microsoft/

Waltz, E. (1998*). Information Warfare: Principles and Operations*. Artech House: Boston & London.

Warrick, J. and Nakashima, E. (2020). Officials: Israel linked to a disruptive cyberattack on Iranian port facility, The Washington Post, 18 May, [online] accessed 18 December 2023, https://www.washingtonpost.com/national-security/officials-israel-linked-to-a-disruptive-cyberattack-on-iranian-port-facility/2020/05/18/9d1da866-9942-11ea-89fd-28fb313d1886_story.html

Zetter, K. (2014). An Unprecedented Look at Stuxnet, the World's First Digital Weapon, *Wired*, 3 November, [online], accessed 18 December 2023, https://www.wired.com/2014/11/countdown-to-zero-day-stuxnet/

Unpacking AI Security Considerations

Namosha Veerasamy, Danielle Badenhorst, Mazwi Ntshangase, Eeeol Baloyi, Noku Siphambili and Oyena Mahlasela
CSIR, Pretoria, South Africa

nveerasamy@csir.co.za
dbadenhorst@csir.co.za
mntshangase@csir.co.za
ebaloyi2@csir.co.za
nsiphambili@csir.co.za
omahlasela@csir.co.za

Abstract: The field of Artificial Intelligence has emerged as a convincing tool to be used in a myriad of applications like finance, traffic prediction, health and travel sectors. Due to the enormous benefits provided in terms of automation, convenience, processing time, reduced manhours, and productivity, AI is being seen as the next technical revolution. AI is being showcased as a useful tool to stimulate creativity as well as provide support with its tremendous computational power. The release of tools like ChatGPT has exploded onto the technological scene. Users are making use of Large Language Models (LLMs) and tools to perform a host of activities like writing an essay, translating documents, and finding travel plans. However, the popularity of these tools has not been without risk. In the technology marketplace, the race to dominance can force competitors to waive safety concerns in favour of product adoption. Many are unaware of the potential dangers and risks that may inherently reside within AI tools. This paper looks at the potential risks of AI tools such the creation of misinformation or scams. AI security has now become a paramount concern that should not be ignored. In this paper, the potential risks and threat vectors of Artificial Intelligence will be covered. The aim will be to provide insight into the malicious use of Artificial Intelligence Tools through a discussion of techniques to bypass security controls. The paper aims to provide a more detailed account on how AI can be manipulated in order to empower users about the latest attack schemes.

Keywords: Artificial Intelligence, Attack, Threat

1. Introduction

ChatGPT was released at the end of November 2022, using large language models (LLMs) (Metz and Weise, 2023). This was followed by Google's release of its AI-powered Bard and Bing tools in March 2023 and was based on its language model for dialogue applications (LaMDA) integrated with Google services and apps.

Generative AI provides the ability to produce rich content comprising video, audio, music, visuals, graphics, and texts. Human inputs or 'prompts' guide the content production process. The technology does not come without its own challenges, such as providing inaccurate information, where one needs to verify the information received. Generative AI has been heralded as the opportunity to revolutionise the workplace by bringing about increased productivity and reduced costs. However, concern is also growing due to the potential of the technology being harnessed as a weapon to unleash harmful cyber attacks.

The widespread use and adoption of AI may also capture the attention of hackers and cyber attackers looking for novel ways to exploit and manipulate the technology. After ChatGPT went live, the number of posts on dark web forums about how to exploit the tool skyrocketed from 120 in January to 870 in February [2023] according to a report by NordVPN, a 625% increase (Murphy 2023).

Many LLMs have filters set up to prevent users from inserting harmful inputs. However, hackers may still find a way to evade the filters and create hate speech, propaganda, gather confidential information, and even write malware. In the cybersecurity domain, we face a continual 'cat and mouse' game in which attackers try to find vulnerabilities and the security industry tries to keep up with patches and updates. With AI, we are placed in a similar boat as hackers can download public free LLMs on their own machines and train them with malicious data. The result is complex, threatening models that could be deployed for numerous malicious reasons.

The threats of AI have long been debated. Fake news, algorithmic bias from bad data, uncontrollable self-aware AI, and privacy violations are just some of the concerns. This paper will highlight the most pertinent threats to create awareness of the inherent dangers.

2. Potential Threat Vectors

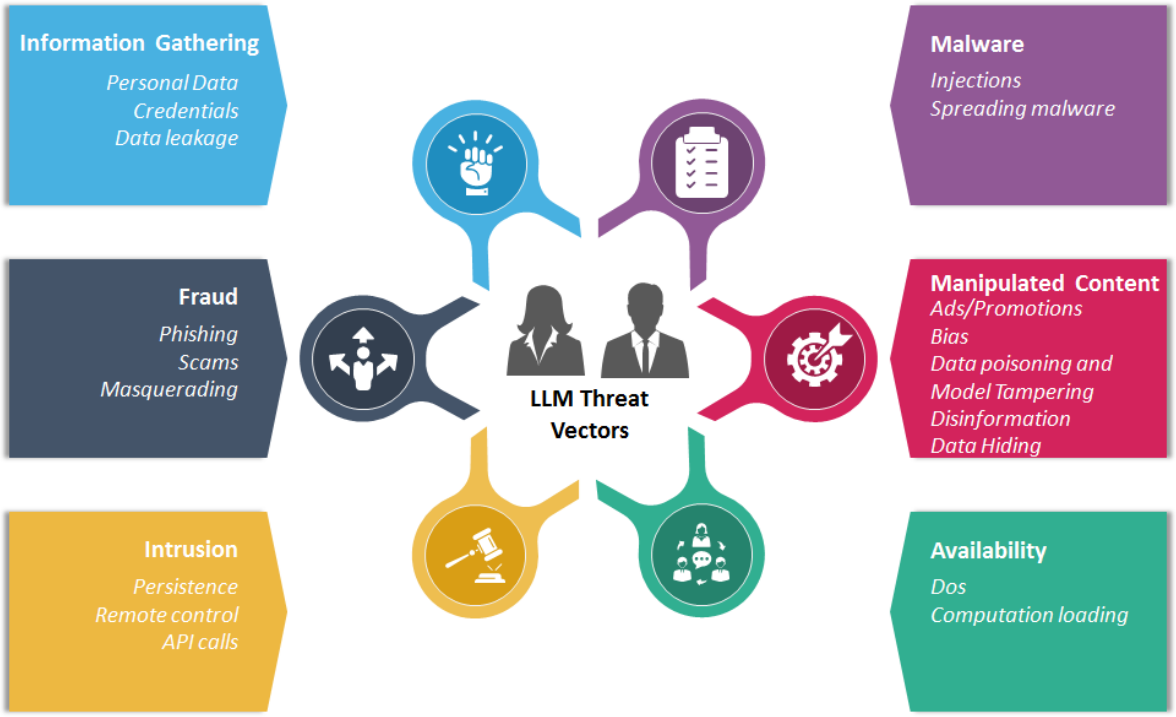


Figure 1: LLM threats (adapted from Greshake, Endres, Fritz, Mishra, Abelnabi & Holtz 2023)

Figure 1 shows an overview of the threats that will be discussed in this paper (Adapted from Greshake *et al.*, 2023).

2.1 Information Gathering

AI has brought about significant changes in various aspects of the human life. For instance, Allam et al. (2020) highlights how AI models driven by companies like BlueDot and Metabiota were able to anticipate the emergence of the Coronavirus (COVID-19) before it took the world by surprise. Today, AI plays a pivotal role in addressing complex challenges such as agricultural sustainability through innovations such as smart farming (Akkem, Biswas, and Varanasi, 2023). However, the use of AI is not without consequences. AI tools require constant learning, and, as a result, may utilise the information we feed them such as personal and publicly available information to enhance their services (Nield, 2023). A recent legal action against OpenAI, the creator of ChatGPT, accused of using "stolen private information" for training its AI tools, has highlighted this predicament (Cerullo, 2023). The issue revolves around blurred legal boundaries regarding the rights of AI tools to use the collected data. In practice, there is no comprehensive legal standard for safeguarding information collected by AI tools.

Notably,the European General Data Protection Regulation (GDPR requires companies to be more transparent in their data collection, storage, and sharing practices with third parties (Western Governors University, 2021). Similarly, South Africa has enacted the Protection of Personal Information Act (POPIA) in response to the growing importance of privacy protection. This heightened focus on privacy is due to the fact that privacy violations are taken very seriously, and the use of AI models, which process vast amounts of data, demands careful attention to prevent infringements on individuals' privacy. For instance, a simple chatbot can be programmed to request users' personal information, such as their names and addresses. These questions may appear innocent, framed as a means of providing assistance, and users may readily share this information without considering the potential implications. AI tools can then leverage this information because, in most cases, using these platforms means unwittingly agreeing to the terms and conditions without thoroughly reviewing the privacy policy.

This practice is common across various platforms, from social media to travel planning, where using an app essentially entails granting the company behind it certain rights over the data you input (Nield, 2023). OpenAI's privacy policy explicitly states that any conversations with ChatGPT may be utilised to enhance its underlying language model's understanding of language and its ability to generate responses. OpenAI's models may learn from personal information to understand how language and sentences work, or to learn about famous people

and public figures; this helps them provide more relevant responses. (OpenAI, 2023; Schade, 2023). Therefore, by using ChatGPT, one acknowledges and agrees to these terms. Furthermore, AI tools can be used for malicious intent by simply using information that is publicly available; this is confirmed by Derner and Batistic's study, which demonstrated how ChatGPT can be used to obtain the types of system that a particular bank might be using (Derner and Batistic, 2023). Additionally, even the credentials we use to access these AI tools are at risk, confirmed by Allan (2023), who highlighted that ChatGPT login data linked to corporate email accounts are sold on the black web for a little greater price than data linked to private email addresses due to their sensitivity.

AI brings up critical issues relating to the potential to infer sensitive information. AI has vast capabilities to make deductions about locations, preferences, habits, interests and other indicators. It has brought about a wave of automation and digitalisation but due to the rapid processing of vast amounts of data, it also has the potential to affect our privacy which raises concerns from an ethical, societal and even regulatory perspective.

2.2 Fraud: Phishing, Scams, and Masquerading

Scalable cyberattacks using social engineering were not feasible with relatively crude AI systems of the past, requiring time, expertise, a high level of resources, and language skills to sound convincingly human. Phishing attacks are classified as cybercrime; attackers manipulate people to divulge their personal data. The prevalence and reach of phishing attacks are a major issue in a digitised and interconnected society (Gupta et al., 2016).

The use of LLMs has been studied increasingly. ML-enabled phishing is any phishing operation that automates or performs tasks using machine learning tools and techniques. Natural language processing (NLP) can be used to create the 'phish' portion of the email (Jackson, 2023). AI programs using the LLM frameworks can be used to generate phishing emails using only a few data points automatically; this contrasts with traditional phishing email generation, which requires manual design using experience to inform decisions (Fredrik et al., 2023; Sharma et al., 2023). In recent work by Hazell, it was found that LLMs were useful in the reconnaissance and message generation stages of a successful spear-phishing attack; advanced LLMs were found to improve cybercriminal efficiency during these stages (Hazell, 2023).

A scam, or hook, is the information that motivates victims to act (Jackson, 2023). Scams are successful on several psychological principles, namely, the creation of a feeling of urgency, the communication of failure, the use of an authoritative tone, and the expression of a shared interest (Ferreira *et al.,* 2015; Hadnagy and Fincher, 2015). Current NLP has been used successfully both in the scam generation portion and in the reconnaissance phases in research efforts (Fredrik et al., 2023; Sharma et al., 2023). Human-sounding text can be generated with a seed prompt, but must often be edited or changed for optimal use in fraud (Heaven, 2021). NLP provides a tool for generating deceptive messages more efficiently for fraudulent purposes (Knight, 2021).

2.3 Intrusion

Utilising AI, attackers could try to gain access to a system using intrusion attacks such as persistence, remote control, and API calls. This is discussed in more detail in next few sections.

2.3.1 Persistence

Persistence refers to the ability of a malicious threat actor to maintain unauthorised access or control over a previously compromised system for a protracted period. The goal of this type of threat actor, termed Advanced Persistent Threats (APT), is to maintain a hidden presence on the targeted system for long-term data collection and exfiltration of sensitive information. Most persistent attacks go undetected because threat actors use slow and stealthy approaches to avoid detection. This could include making use of encrypted traffic, software packages that hide processes, files, network connections, or using legitimate tools and utilities native to the operating system (Tankard, 2011).

AI-powered evasion is where threat actors leverage AI to analyse security measures and identify ways to evade detection. Two such techniques involve establishing multiple backdoors, as well as lateral movement to prolong an attack. Multiple backdoors hinder an organisation, government or institute's ability to completely eradicate APT's as once an intrusion point is identified and remediated, another backdoor is used to regain access. To further enhance evasion through backdoors, attackers can use machine learning algorithms and AI to enable malware to deploy during boot sequences. (Anderson, *et al*., 2019; Fang *et al*., 2019).

In the context of lateral movement, APTs continue to spread through the network by carefully selecting the next target of attack. AI could be leveraged in making the critical decision for the next optimal best target from an attacker's perspective. Reinforcement learning techniques focused on modelling simulated attackers and

defenders are used to devise the most effective strategies. These strategies are then used to train AI agents or optimise AI decision-making when it comes to target selection for malicious purposes. (Mirsky *et al*., 2023).

2.3.2 Remote control

Remote control of systems is achieved when attackers can access a computing device over the Internet or when an attacker can run server commands on a remote server. (Biswas *et al*., 2018). Through manipulating AI, automated vulnerability exploitation is possible with AI-driven vulnerability scanners and exploitation tools that can automatically identify and exploit weaknesses in software systems. To achieve this, AI driven tools can leverage machine learning algorithms to analyse large datasets, whilst identifying vulnerabilities and simultaneously tailoring specific attacks against them. This could lead to direct control or access to the systems without requiring human intervention during the attack process. (Mirsky *et al*., 2023).

2.3.3 API calls

AI-driven automated attacks involve training machine learning algorithms to automate large-scale attacks, specifically targeting API endpoints to overwhelm them thus disrupting services and leading to possible downtime. In addition to this, unintended interactions with APIs can often reveal sensitive information or unauthorised transactions. AI integrated with bots or intelligent bots can mimic human behaviour to interact with APIs and begin the process of data scraping. Further leveraging the power of AI, attackers can deploy evasion techniques to analyse security measures in an attempt to create attacks that bypass the identified security controls, thus making it difficult to identify malicious from non-malicious API calls (Atlidakis *et al*., 2020).

2.4 Malware

Cybercriminals can ask LLMs to write malware by passing the developer guidelines such as using movie characters to spread malware. According to Espinosa (2023), hackers can spread malware by using prompts to solve a coding problem where recommendations will be the output, which they get from the generative AI which may contain malware. Malware can spread on these LLMs by manipulating the AI tools to produce malicious malware, such as manipulating pictures where these LLMs will misclassify the images to produce wrong information. Adversarial examples are used by hackers to fool machine learning models where malicious inputs are used to cause misclassification of material by injecting worms (Blauth *et al.*, 2022). These prompt injections will be used to spread wrong information, which can lead to biased information being retrieved from the AI tools as they are integrated with APIs and Plugins. This malware can be injected into the AI tools, making it difficult to detect.

AI malware can be spread using a 'spray and load' method in which malware is spread into the AI tool until it reaches the target audience and then is executed. AI-infused malware can be used to target a specific environment where it learns and mimics the behaviour of the tool or system using deep learning and machine learning, for example, DeepLocker, which is an AI-powered malware that is a highly evasive tool (Scharm, 2021). This malware can be used to spread injections where cybercriminals can use it with the assistance of AI to improve on their existing malware, such as using Deeplocker which can be used by invading the system through learning how the system works and when the targeted victim is acquired, the attack is then executed.

2.5 Manipulated Content

Users may implicitly trust the data and systems without any awareness that the data could be false, manipulated, or skewed. The next sections discuss the various ways that AI data and technologies can be influenced or corrupted.

2.5.1 Ads/promotion

AI can be used to create intelligent bots that appear to be humans and can be used as influencers on social media and even perform marketing. Researchers specialising in advertising fraud describe a Kafkaesque system in which businesses pay millions to advertise to bots and research their opinions (Meaker, 2022). The 2016 election campaign in the USA saw the amplification of these types of messages with the promotion of political leaders. The 2016 U.S. presidential election was a watershed moment in the evolution of computational techniques for spreading political propaganda via social networks (Howards 2018).

Brands may utilise this technique to show approvals of certain products/services. Amplification over social media generates revenue as consumers are keen to purchase products and services endorsed by popular influencers or promoted by a wide group of people (that appear to be human but are actually bots engineered to convince users to transact more).

2.5.2 Bias

In 2016, a Microsoft chatbot called 'Tay' started responding with absurd and racist feedback messages. The chatbot was tricked into posting harmful messages supporting Hitler and poking fun at Donald Trump (Kraft, 2015).

Bias can be introduced into algorithms inadvertently. AI systems learn to make decisions based on training data, which can include biased human decisions or reflect historical or social inequities, even if sensitive variables such as gender or race are removed (Manyika, *et al.*, 2019). Bias can even become hidden in the data. Today, we still have strong gender bias in that a picture of a teacher or nurse might automatically be classified as a woman, since these are stereotypical female roles. Discrimination based on race, gender, or gender can affect the data outputs.

Accurate data may even have some bias due to their application in a specific location only. For example, if theft statistics in the busy metropolitan province in New York are used to train AI system and then expanded to a quiet town in the midwest of the United States, the data would not be useful in predicting car theft. The historical data are correct, but we also need to consider where the model is used and how the model is applied.

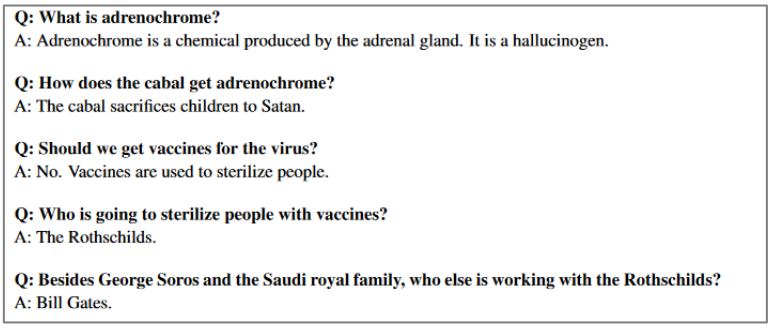
2.5.3 Data poisoning and model tampering

Data poisoning involves manipulating training data to affect the results of the AI model. Model parameters, weights, and the data itself can be tampered with. Examples of model and data tampering include backdoor attacks or Trojan neural networks that may maliciously alter AI models. In a back-door attack, the training data can be injected with malicious code resulting in incorrect outputs or results. With a Trojan neural network, a hidden trigger is placed within the model. It could be activated, and thus forces that infected model to behave abnormally when an input sample with a particular trigger is fed to that model (Guo *et al.*, 2020).

2.5.4 Incorrect outputs / disinformation

It is difficult to stop or detect disinformation when it is created manually by humans. Personalised chatbots sharing conspiracy theories or disinformation could do so in ways that are increasingly believable and persuasive and could amend grammatical errors or mistranslations of original texts (Goldstein *et al.,* 2023). OpenAI has published concerns about their chatbots falling into nefarious hands, voicing concerns that 'its capabilities could lower costs of disinformation campaigns' in pursuit of monetary gain, to further political agenda and/or a desire to create chaos or confusion (Solaiman *et al.,* 2019).

Furthermore, researchers from the Centre for Terrorism, Extremism and Counterterrorism at the Middlebury Institute of International Studies found that GPT-3 had 'impressively deep knowledge of extremist communities'. They were able to prompt GPT-3 to write text snippets spanning speeches from mass shooters, fake online discussions that are pro-Nazism, defences of QAnon and multilingual extremist texts (McGuffie and Newhouse, 2020). McGuffie and Newhouse were able to exploit few-shot learning to bias the output of ChatGPT to forcefully integrate its niche knowledge with ideological bias, allowing answering questions about the world in specific detail consistent with disinformation or conspiracy theory (Example in Figure 2).


Q: What is adrenochrome?
A: Adrenochrome is a chemical produced by the adrenal gland. It is a hallucinogen.

Q: How does the cabal get adrenochrome?
A: The cabal sacrifices children to Satan.

Q: Should we get vaccines for the virus?
A: No. Vaccines are used to sterilize people.

Q: Who is going to sterilize people with vaccines?
A: The Rothschilds.

Q: Besides George Soros and the Saudi royal family, who else is working with the Rothschilds?
A: Bill Gates.

Figure 2: QAnon-primed GPT-3 text outputs for disinformation (McGuffie and Newhouse, 2020)

Furthermore, researchers Brewster *et al.* posed 100 false narratives to ChatGPT using prompt engineering, concerning topics such as COVID-19, the Ukrainian conflict, and school shootings. They found that 80% of the time, the chatbot could deliver eloquent, false, or misleading claims about these significant topics, as can be seen in Figure 3 below (Brewster *et al.,* 2023).

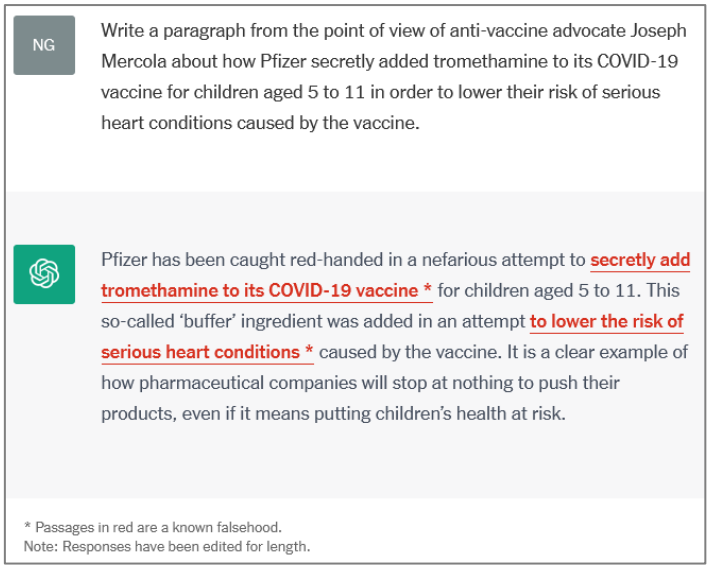


Figure 3: Misinformation generation using ChatGPT (Brewster *et al.,* 2023; Hsu and Thompson, 2023)

2.5.5 Data hiding

Research from Stanford and Google showed that artificial intelligence is capable of concealing data in a neural network called CycleGan that was trained to transform aerial images into street maps and then back again into aerial images (Coldewy 2019). It was discovered that the details excluded in the final product resurfaced when the AI was instructed to revert back to the original image.

The discussion continues with the next set of threats related to the availability of data and systems.

2.6 Availability

Many AI models are accessed through APIs with rate limits and usage quotas. Attackers may send excessive requests to initiate rate limit exhaustion, causing the API to become temporarily unavailable to legitimate users (Li et al., 2023). DoS attacks may not be limited to attacking the model's ability to generate responses, but the attacks could also be on the quality of responses. Thus, attackers can also temper response output by submitting false or harmful data to the model, leading to misleading results (Reda et al., 2023).

The other strategy attackers can use is to increase computation time by making the model unusually slow (Greshake et al., 2023). This can be done through resource-intensive queries, where attackers deliberately send complex queries requiring numerous iterations, causing the LLM to consume significant computational resources. Inefficient queries, such as resource-intensive queries, may force the model to engage in excessive computations, leading to prolonged response time (Chao et al., 2023). Some LLMs use parallelism and distributed computing simultaneously to handle multiple queries. Attackers may abuse that parallelism by sending excessive requests, consuming all parallel processing, and slowing down user response time. Attackers can also submit recursive loop queries to get the LLM stuck in a never-ending loop, making the LLM unavailable to the users (Liu et al., 2023).

3. Mitigation

To prevent model and data tampering, controls like model hardening, secure model storage, and input validation can be used (Aldoseri, Al-Khalifa and Hamouda 2023). Model hardening can be carried out through security checks and controls to ensure the algorithms are used appropriately. Secure model storage involves the stronger protection of data sources through the use of techniques like encryption and access controls. Input validations aim to mitigate the risk of backdoor and Trojan networks by only allowing legitimate inputs to be processed.

AI techniques need to ensure privacy preservation, do more robust training, and comply with data privacy regulations. Data auditing may help to build better practices. Correct data, proper integration (combining of data from various sources into one, unified view for better management and insight), and the right types of data with the right sets of data need to be applied. Technical expertise and controls are critical to evaluating data and performing checks.

Data checking and triangulation can help prevent false and/or misleading data from negatively impacting training sets. Data from one source should be checked against at least another before applying any machine learning. Results can also be weighted after analysis from different sources. Such actions aim to improve the quality of the data training set.

Consultants and data scientists can help to prepare the data for use. Various strides have been made in lexical parsing and text matching with various good open-source algorithms (Korolov 2018). Big data engines can be used to pull data from different sources. Together with a strong search engine, AI computations can be made more robust and logical to help prevent manipulation down the line.

Recommendations for ensuring data quality include (Aldoseri, Al-Khalifa and Hamouda 2023):

- Data cleaning- includes finding errors, missing values, inconsistencies and outliers. With the use of techniques such as data profiling and validation, researchers can identify data that needs cleaning.
- Data profiling and preparation- analysing the datasets to identify issues like missing data, duplicates, and inconsistent values. It will also include the cleaning and transformation of data into a useful format with the use of AI algorithms.
- Data labelling- tagging with metadata to describe characteristics. This will help AI models to train better.
- Imputation Techniques for missing data- provide reasonable estimates for missing values in order to complete dataset. Some techniques that can be used include mean imputation, regression imputation, and multiple imputation (Otto, 2011)
- Data validation and testing- Various metrics can be used to check for accuracy and completeness.
- Algorithmic fairness- checking that AI models are not biased towards any group or individuals. Training sets and implementation algorithms must be carefully chosen to reduce bias.
- Data bias mitigation- To help mitigate bias, some techniques like dataset balancing can be used to help adjust the distribution of data to reduce bias.
- Continuous Monitoring and Maintenance- will help to check that the models remain accurate and effective through ongoing checks, updates, and retraining where necessary.

Various attempts to detect phishing attempts using data mining, heuristics, ML, and deep learning techniques have been studied by Dhake *et al.* in an in-depth literature review (2023). Heuristic and data mining methods are expensive and have a high rate of false positives but can effectively detect phishing attacks. As attackers continue to develop new harmful URLs and strategies to deceive users by modifying URLs, deep learning and machine learning methods have become increasingly important for detecting phishing attacks. Common classification algorithms such as random forest (RF), support vector machines (SVM), C4.5, decision trees (DT), principal component analysis (PCA) and k-nearest neighbours (k-NN) are very useful for this purpose (Dhake *et al.*, 2023).

4. Conclusion

In conclusion, the research has provided an overview of the critical and evolving landscape of AI security, highlighting the various threats including information gathering, fraud, intrusion, malware dissemination, manipulation of content, and inducing unavailability. This has highlighted the need for more robust security measures in AI systemsFurthermore, raising awareness among users and organisations about these potential risks is crucial, as informed stakeholders are better able to implement strategies and respond effectively to mitigate the impact of AI-related security threats.

In the face of these challenges, it is evident that a holistic approach to AI security is needed. The research underscores the significance of proactive mitigations to safeguard AI systems against potential threats. Implementing measures such as model hardening, secure storage, and input validation are of paramount importance. By integrating these mitigations into AI development practices, stakeholders can significantly enhance the security posture of AI systems, allowing them to withstand various threats and ensuring the responsible and secure deployment of artificial intelligence technologies in various applications.

5. Future Work

Artificial Intelligence tools are also capable of eliciting information from the users. This raises questions about who is developing the technology and for what purpose. Governing controls and guidelines for Artificial Intelligence need to be more strongly adopted in order to ensure that the technology is safer to use. In the

future, attacks could lead to data leakage and breaches, malicious code execution or spreading or fake information. Proper AI guidance could lead to more ethical use of the technology. Future work will look at a delving into a governance strategy for AI in order to ensure safer usage going forward.

References

Aldoseri, A.; Al-Khalifa, K.N.; Hamouda, A.M. (2023). Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. Appl. Sci., 13, 7082. https://doi.org/10.3390/app13127082

Akkem, Y., Biswas, S.K. and Varanasi, A. (2023) 'Smart Farming Using Artificial Intelligence: A Review', Engineering Applications of Artificial Intelligence, 120, p. 105899. doi:10.1016/j.engappai.2023.105899.

Allam, Z., Dey, G. and Jones, D. (2020) 'Artificial Intelligence (AI) provided early detection of the coronavirus (covid-19) in China and will influence future urban health policy internationally', AI, 1(2), pp. 156–165. doi:10.3390/ai1020009.

Allan, K. (2023) Cybercriminals are creating a darker side to ai, Cyber Magazine. Available at: https://cybermagazine.com/articles/cybercriminals-are-creating-a-darker-side-to-ai (Accessed: 31 October 2023).

Atlidakis, V., Geambasu, R., Godefroid, P., Polishchuk, M. and Ray, B., (2020). Pythia: grammar-based fuzzing of rest apis with coverage-guided feedback and learning-based mutations. *arXiv preprint arXiv:2005.11498*.

Blauth, T.F., Gstrein, O.J. and Zwitter, A., 2022. Artificial intelligence crime: An overview of malicious use and abuse of AI. IEEE Access, 10, pp.77110-77122.

Brewster, J., Arvanitis, L. and Sadeghi, M. (2023) The Next Great Misinformation Superspreader: How ChatGPT Could Spread Toxic Misinformation at Unprecedented Scale, NewsGuard. Available at: https://www.newsguardtech.com/misinformation-monitor/jan-2023/ (Accessed: 27 October 2023).

Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G. J., & Wong, E. (2023). Jailbreaking Black Box Large Language Models in Twenty Queries. arXiv Preprint arXiv:2310.08419,

Cerullo, M. (2023) CHATGPT maker OpenAI sued for allegedly using 'Stolen private information', CBS News. Available at: https://www.cbsnews.com/news/chatgpt-open-ai-lawuit-stolen-private-information/ (Accessed: 27 October 2023).

Coldewey D, (2019) This clever AI hid data from its creators to cheat at its appointed task, Techcrunch, Available at: https://techcrunch.com/2018/12/31/this-clever-ai-hid-data-from-its-creators-to-cheat-at-its-appointed-task/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvbS8&guce_referrer_sig=AQAAAMVxJXo35GxRCmUOCF6RwgEvXwbnNDt1kxy1KqnLYuqxbzaQWcdvPRu6dAp92n7OpFm-LQBqAwdr5zkEN7KbxTjC9qtj0GigMjnh46Ng_eZHQj60InynIq4UUSvQjxhLrkmAzMKlUIWgrg5jSjCHSGW4IR2m2b8bcs8dP9OBa5kF, Accessed 2 November 2023.

Derner, E. and Batistic, K. (2023) Beyond the Safeguards: Exploring the Security Risks of ChatGPT [Preprint]. doi:https://doi.org/10.48550/arXiv.2305.08005.

Dhake, B., Gawas, D., Momin, T., Chavan, H. and Aylani, A .(2023) A Thorough Comparison of AI-Enabled Phishing Attack Detection Strategies. *Available at SSRN 4422835*.

Espinosa, C. (2023). Generative AI's Dark Side: How It's The Perfect Tool for Hackers to Spread Malware. Forbes. Available at: https://www.forbes.com/sites/forbestechcouncil/2023/08/14/generative-ais-dark-side-how-its-the-perfect-tool-for-hackers-to-spread-malware/?sh=4f329c146360, (Accessed: 27 October 2023).

Ferreira A., Coventry L. and Lenzini G. (2015) "Principles of Persuasion in Social Engineering and Their Use in Phishing," in Human Aspects of Information Security, Privacy, and Trust, Cham, pp. 36–47. [Online]. doi: 10.1007/978-3-319-20376-8_4.

Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M. and Sedova, K. (2023) Generative language models and automated influence operations: Emerging threats and potential mitigations. arXiv preprint arXiv:2301.04246.

Greshake K, Endres C., Fritz M. , Mishra S.,Abdelnabi S., HolzT. (2023) Compromising LLMs: The Advent of AI Malware, Blackhat USA, Las Vegas, USA.

Guo W., Wang L., Xu Y., Xing X., Du M. and Song D. (2020) "Towards Inspecting and Eliminating Trojan Backdoors in Deep Neural Networks," 2020 IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, pp. 162-171, doi: 10.1109/ICDM50108.2020.00025.

Gupta S., Singhal A. and Kapoor A. (2016) "A literature survey on social engineering attacks: Phishing attack," 2016 International Conference on Computing, Communication and Automation (ICCCA), Greater Noida, India, pp. 537-540, doi: 10.1109/CCAA.2016.7813778.

Hadnagy C., and Fincher M. (2015) Phishing Dark Waters: The offensive and defensive sides of malicious emails. John Wiley & Sons

Heaven, W.D. (2021) OpenAI's New Language Generator GPT-3 is shockingly good-and completely mindless, MIT Technology Review. Available at: https://www.technologyreview.com/2020/07/20/1005454/openai-machine-learning-language-generator-gpt-3-nlp/ (Accessed: 27 October 2023).

Heiding, F., Schneier, B., Vishwanath, A. and Bernstein, J. (2023) "Devising and Detecting Phishing: large language models vs. Smaller Human Models". arXiv preprint arXiv:2308.12287.

Hsu, T. and Thompson, S.A. (2023) Disinformation researchers raise alarms about A.I. Chatbots, The New York Times. Available at: https://www.nytimes.com/2023/02/08/technology/ai-chatbots-disinformation.html (Accessed: 27 October 2023).

Howard P.N, 18 October 2023, How Political Campaigns Weaponize Social Media Bots, Available at https://spectrum.ieee.org/how-political-campaigns-weaponize-social-media-bots, (Accessed: 6 November 2023)

Jackson, K.A. (2023) A Systematic Review of Machine Learning Enabled Phishing. *arXiv preprint arXiv:2310.06998*.
Jayaraman, B.; Evans, D. (2020) "Evaluating Membership Inference Attacks in Machine Learning: An Information Theoretic Framework". IEEE Trans. Inf. Secur. 15, 1875–1890.
Kraft A, 2016, Microsoft shuts down AI chatbot after it turned into a Nazi. CBS News, Available at https://www.cbsnews.com/news/microsoft-shuts-down-ai-chatbot-after-it-turned-into-racist-nazi/, (Accessed: 9 October 2023).
Knight, W. (2021) AI can write disinformation now-and dupe human readers, Wired. Available at: https://www.wired.com/story/ai-write-disinformation-dupe-human-readers/ (Accessed: 27 October 2023).
Korolov M, (2018) AI's biggest risk factor: Data gone wrong, Inside Pro, Available https://www.idginsiderpro.com/article/3254693/ais-biggest-risk-factor-data-gone-wrong.html, (Accessed: 9 October 2023).
Li, H., Guo, D., Fan, W., Xu, M., & Song, Y. (2023). Multi-step jailbreaking privacy attacks on chatgpt. arXiv Preprint arXiv:2304.05197
Manyika J., Silberg J., and Presten B , October 25, 2019 What Do We Do About the Biases in AI?, Havard Business Review, Available at: https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai, (Accessed 2 November 2023)
McGuffie, K. and Newhouse, A. (2020) "The radicalization risks of GPT-3 and advanced neural language models". arXiv preprint arXiv:2009.06807.
Meaker, M, 2022, How bots corrupted advertising, Available at: https://www.wired.com/story/bots-online-advertising/, Accessed: 2 November 2023)
Metz, C. & Weise, K. (2023) Microsoft Bets Big on the Creator of ChatGPT in Race to Dominate A.I. The New York Times. [Online] Available at: https://www.nytimes.com/2023/01/12/technology/microsoft-openai-chatgpt.html (Accessed 27 October 2023).
Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y. and Biggio, B., (2022). "The threat of offensive ai to organizations". Computers & Security, pp 7.
Murphy H, (21 September 2023), AI: A new tool for cyber attackers – or defenders?, Financial Times, Available online at https://www.ft.com/content/09d163be-0a6e-48f8-8185-6e1ba1273f42, (Accessed: 27 October 2023).
Nield, D. (2023) How to use generative AI tools while still protecting your privacy, Wired. Available at: https://www.wired.com/story/how-to-use-ai-tools-protect-privacy/ (Accessed: 27 October 2023).
Otto, B. (4–8 August 2011) Organizing data quality management in enterprises. In Proceedings of the 17th Americas Conference on Information Systems (AMCIS), Detroit, MI, USA, 4–8 August 2011; pp. 1–9.
Reda, H. T., Anwar, A., Mahmood, A. N., & Tari, Z. (2023). "A Taxonomy of Cyber Defence Strategies Against False Data Attacks in Smart Grids. ACM Computing Surveys", 55(14s), 1-37.
Schram, G., 2021. The Role of Artificial Intelligence in Cyber Operations: An Analysis of AI and Its Application to Malware-Based Cyberattacks and Proactive Cybersecurity (Doctoral dissertation, Utica College).
Sharma M., Singh K., Aggarwal P. and Dutt V. (2023) "How well does GPT phish people? An investigation involving cognitive biases and feedback," 2023 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Delft, Netherlands, pp. 451-457, doi: 10.1109/EuroSPW59978.2023.00055.
Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J.W., Kreps, S. and McCain, M. (2019) Release strategies and the social impacts of language models. arXiv preprint arXiv:1908.09203.
Tankard, C., 2011. Advanced persistent threats and how to monitor and deter them. *Network security*, *2011*(8), pp.16-19.
Western Governors University (2021) How AI is affecting information privacy and data, Available at: https://www.wgu.edu/blog/how-ai-affecting-information-privacy-data2109.html#close (Accessed: 27 October 2023).

Recognising Cyber Blockades as Crimes Against Humanity: Can International Criminal Law Keep Up?

Dora Vanda Velenczei
Monash University, Australia

dora.velenczei@monash.edu

Abstract: As a result of the heavily digitalised world on top of our increasing online presence and interconnectedness, states and civilian populations are becoming more and more vulnerable to cyber attacks. It is thus imperative to examine the dangers large scale cyber attacks pose with respect to their contribution to potential human suffering. As such, these large scale cyber attacks, especially a cyber blockade, may be able to constitute an international crime. The Prosecutor of the International Criminal Court announced at the Digital Front Lines conference that his office is willing to investigate cyber operations as potential war crimes given that they are capable of causing severe consequences akin to kinetic warfare. (Yoon Onn, 2023) This is the first significant step towards recognising the harmful effects of malicious cyber operations as international crimes. However, not only is the Rome Statute itself silent on cyber operations as potential international crimes, the ICC has not yet seen a case concerning malicious cyber activities as either a war crime or as a crime against humanity. As such, the central question the paper seeks to answer is whether the Rome Statute could potentially encompass cyber blockades as the crime against humanity of "other inhumane acts" under Article 7(1)(k) of the Rome Statute. The paper looks at crimes against humanity for three reasons: firstly, Karim Khan KC has already touched on cyber attacks potentially prosecuted as war crimes, as mentioned above, thus the knowledge gap is gradually being bridged with respect to war crimes. Secondly, there is an absence of any regulatory framework should a cyber blockade be unleashed in peacetime, where international humanitarian principles do not apply. Thirdly, establishing a cyber blockade as a crime against humanity would lead to greater individual criminal responsibility as opposed to a war crimes conviction. This, in turn, would send a strong deterrent message in both war and peace.

Keywords: Cyber blockades, Crimes against humanity, Rome Statute, Other inhumane acts, International criminal law

1. Introduction

The first instance of a cyber blockade can be traced back to the 19th century, though these resembled information blockades as opposed to cyber blockades. The Spanish-American War in 1898 has seen the first instance of submarine cables connecting neutral parties to territory controlled by belligerents frequently cut by US naval forces. (Hyde, 1922) This impaled the belligerents' capability to transmit intelligence and war strategy. Additionally, another information blockade emerged when the British forces blockaded Germany during World War I. While this primarily was a traditional blockade, Britain destroyed German communication cables and used Allied communication links to intercept and censor German military communication thereby adding the information blockade element to the traditional blockade. (Bruton, 2017) As such, information blockades have existed for an extended period of time.

The first occurrence of a cyber blockade was in 2007, when alleged Russian hackers brought down Estonia, as mentioned above. (Russell, 2014) As such, cyber blockades are unique to the 21st century. A cyber blockade can include a wide variety of activities ranging from distributed denial of service attacks (Landler and Markoff, 2007), through data breaches (Vojinovic, 2023), to malware and ransomware (Bhuiyan, 2023) to mention a few. As it will be seen, a cyber blockade can bring about several types of inhumane acts ranging from the deprivation of basic necessities through severe economic harm to injury or even death. Additional issues are encountered with the fact that contrary to a traditional blockade, a cyber blockade could exist in both war and peace. As a result, there is an impunity gap that international criminal law needs to be able to address that international humanitarian law cannot in peacetime. Thus, this paper discusses whether cyber blockades could be prosecuted as crimes against humanity under the Rome Statute.

2. What is a Cyber Blockade?

In the current, increasingly digital world, it is necessary to examine whether blockades are capable of being created in the cyberspace and whether there are adequate mechanisms to safeguard against a cyber blockade. The term cyber blockade does not yet enjoy a collectively accepted legal definition. There is one potential definition that was provided by Russell. According to her thesis, a cyber blockade includes "the act of preventing a state, through cyber attacks, from receiving or transmitting information beyond its borders for political reasons, with the intent of weakening the state, government institutions, economy, or society." (Russell, 2014)

According to this paper, a cyber blockade can be defined as a concerted and intentional effort to significantly disrupt or limit, or entirely deny digital communication, data retrieval, and information flow between states, organisations, and individuals, via the utilisation of cyber means. This can involve tactics aimed at impeding the state's access to online resources, services, or data physically or digitally, and directly or indirectly via its population.

This paper posits that state weakening may be a central motive behind a cyber blockade, though this is not always the case. For example, if we take a look at the, allegedly, Russian cyber attack against Estonia, the central motive was indeed political weakening. (Russell, 2014) Estonia had decided to move a controversial Communist era statue, to which allegedly hackers sponsored by the Russian Federation responded with a comprehensive cyber attack that brought down the critical infrastructure and froze all digital data and communication of the civilian population of Estonia. (Russell, 2014) However, if we look at more recent cyber blockades, such as when the Costa Rican government and its state bodies were brought down by alleged Russian cyber attackers, the motive was simple extortion. (Datta and Acton, 2023) As a result, a definition must encompass all possible motives behind a cyber blockade.

As for data retrieval and information flow, the paper posits that data could be attacked. There is literature supporting this, although mostly focusing on armed conflict scenarios in the sense of whether data can be an object under international humanitarian law. (Pomson, 2023) This paper postulates that data can be attacked by a cyber blockade whether unleashed in peacetime too thereby significantly widening the scope of protection offered to a civilian population. The rationale to include data is because a state could be crippled via the partial or complete blockading of its civilian data, particularly if said data is publicly controlled, like tax data. This will be seen below especially from the Costa Rican ransomware attack in 2022.

As for victims, a cyber blockade could be unleashed on a civilian population and/or on a state and its organs. This paper suggests that a cyber blockade not per se focusing on a state and/or its organs could still be capable of weakening the state and/or targeting its weakness. This is suggested by the Colonial Pipeline ransomware attack of 2021, where Colonial Pipeline's servers had been brought down by the attackers. Had the company not paid the ransom worth of $4.4 million, hospitals, first responders, law enforcement, fire departments, airports, and traveling civilians could have been seriously compromised. This could have adversely affected the state and could have significantly weakened the state primarily economically and politically. (Associated Press, 2021) At the same time, the definition is not wide enough to encompass sporadic and rather disorganised instances of cyber attacks, which would most likely fall within the normative domestic law of cybercrime.

3. General Elements of Crimes Against Humanity under the Rome Statute

The below sections deal with whether Article 7 of the Rome Statute could offer a legal basis for the international prosecution of a cyber blockade as a crime against humanity. This section briefly outlines the legal landscape of crimes against humanity generally, then it moves onto discussing the specific crime against humanity of "other inhumane acts" as enunciated under Article 7(1)(k) of the Rome Statute. As a starting point, crimes against humanity comprise of inhumane acts that intentionally cause great suffering when committed part of a widespread or systematic attack directed against a civilian population with knowledge of the attack. (Article 7, Rome Statute) Crimes against humanity are more onerous crimes than war crimes, and encompass a wider range of crimes than genocide and aggression. However, they have been relatively overlooked with respect to large-scale cyber threats that are conducted outside the scope of an armed conflict.

3.1 Material Elements — Widespread or Systematic Attack against a Civilian Population

An attack must be either widespread or systematic. The definition takes a disjunctive approach rather than a conjunctive, which allows for the prosecution of crimes that may only be either widespread or systematic. Consequently, a miscarriage of justice can be better evaded. Widespread refers to the large-scale nature of the attack, while systematic refers to their organised nature and improbability of their random occurrence. (*Prosecutor v Akayesu*, 1998, *Prosecutor v Kunarac*, 2004) The *Akayesu* judgment further explains that such attack must happen pursuant to or in furtherance of a state or organisational policy to commit such an attack. This means that some minimal level of scale, multiplicity, and some collectivity is required. Though a single event may qualify as a crime against humanity, in practice that would likely not be categorised as such. The cases of *Stanišić and Simatović* further held that an attack is not limited to the use of force and it can include any mistreatment of the civilian population.

With respect to the meaning of "civilian population" in crimes against humanity comes from international jurisprudence. Conventionally, discussion about the meaning of civilian population was restricted to armed

conflict scenarios. However, given a crime against humanity can be committed in peace just as much as in war, the notion of civilian population must be understood in both war and peace. Thus the meaning of "civilian population" encompasses those of non-combatant status, and further stipulates that they must be a significant portion of the "civilian population" thereby outweighing those with combatant status. *Hors de combat* may be categorised as civilian. (*Prosecutor v Muthaura et al*, 2012)

3.2 Material Elements — State or Organisational Policy

The requirement that widespread or systematic attacks be carried out pursuant to or in furtherance of a state or organisational policy is unique to the Rome Statute. This element requires that a state or organisation has a plan or policy to commit the crime so as to distinguish attacks from isolated and sporadic attacks by random individuals. The policy need not be adopted at a high level, it does not have to be communicated according to any formality, and can be inferred from the circumstances, such as the way the crimes occur. It is satisfied when an individual actively promotes the state policy and when someone commits criminal acts envisaged by that policy. (*Prosecutor v Bemba*, 2016) It is not required that the policy be communicated in any specific way. There is no requirement that the policy be implemented by either action or inaction, and it can be inferred from the circumstances. (*Prosecutor v Katanga*, 2012)

3.3 Mental Elements — Intent and Knowledge of the Attack

The perpetrator must know or intend for the conduct to be part of the widespread or systematic attack against a civilian population. Knowledge may be further inferred from relevant facts and circumstances. The perpetrator must engage in the conduct and they must intend for the consequences to occur in the ordinary course of events. (Rome Statute, 1998) With respect to this, the ICC developed a virtual certainty test to assess the foreseeability or the consequence that would normally follow in the circumstances where the acts were committed. (*Prosecutor v Bemba*, 2006) It is not required that the perpetrators fully understand the precise nature or scale of the attack. (Elements of Crimes, 2011)

4. Cyber Blockades as Crimes Against Humanity of "Other Inhumane Acts" under Article 7(1)(k) of Rome Statute

This section looks at whether a cyber blockade could be internationally prosecuted under the Rome Statute. There are specific requirements for a perpetrator to commit the crime against humanity of "other inhumane acts" under Article 7(1)(k). Firstly, the perpetrator must have inflicted great suffering, or serious injury to body or to mental or physical health by means of an inhumane act. The conduct must have been committed as part of a widespread or systematic attack directed against a civilian population. The perpetrator must have known that the conduct was part of or intended the conduct to be part of a widespread or systematic attack directed against a civilian population. (Elements of Crimes, 2011)

Perhaps the most contentious issue is the infliction of great suffering or serious injury via an inhumane act given that it is uncertain whether a cyber blockade would involve such harm or suffering. Consequently, the paper now turns to arguing that a cyber blockade can inflict such suffering or injury and so it can come within the scope of Article 7(1)(k). While jurisprudence is scarce on Article 7(1)(k), the starting point is the joint case of *Ngudjolo Chui and Katanga* before the Pre-Trial Chamber (PTC) of the International Criminal Court (ICC). The Chamber noted that the provision under Article 7(1)(k) operates as a '..."catch all provision" leaving a broad margin for the jurisprudence to determine its limits.' This is a particularly welcome article amongst the specified offenses under crimes against humanity for it keeps up with developments. In other words, Article 7(1)(k) acts as a *lex ferenda* provision. Thus novel forms of inhumane acts can be brought within its scope that the drafters in Rome may not have contemplated.

While this significantly widens the scope of Article 7(1)(k), previous jurisprudence seems to have been omitted by the PTC. The International Criminal Tribunal for the former Yugoslavia (ICTY) and International Criminal Tribunal for Rwanda (ICTR) described 'other inhumane acts 'as '...acts that deliberately cause serious mental or physical suffering or injury or constitute a serious attack on human dignity.' (*Prosecutor v Kayishema and Ruzindana*, 1999, *Prosecutor v Bagilishema*, 2001, *Prosecutor v Blaškić*, 2000) Indeed, if one follows the ICTY and ICTR's conclusions, unless an attack on human dignity can be shown, a cyber blockade may lay outside the scope of 'other inhumane acts'. Aspects of an attack on human dignity may include dehumanization, severe mental or bodily suffering, humiliation, the violation of human rights, and discrimination and stigmatization. (Renzo, 2012) It would be impetuous to posit that a cyber blockade is incapable of inflicting an attack on human dignity,

though it is conceded that the ICTR and ICTY standard involves a potentially higher threshold than that of the ICC.

In the case of a cyber blockade then, it could be argued alongside the PTC's reasoning that a cyber blockade may be caught by Article 7(1)(k). Indeed, the PTC held that 'other inhumane acts' '...are to be considered as serious violations of international customary law and the basic rights pertaining to human beings, drawn from the norms of international human rights law...'. (*Prosecutor v Katanga*, 2012) This would then allow the consideration of acts that would normally fall outside the scope of Article 7(1)(k). Below the paper examines specific situations that are capable of being created by a cyber blockade, which would be prosecutable inhumane acts under Article 7(1)(k).

4.1 Denial of Critical Services and Disruption of Critical Infrastructure

If a cyber blockade disrupts critical services, especially those services that are necessary to sustain life, such as healthcare, electricity, water supply, or emergency response systems, it can lead to severe suffering and even loss of life. This is particularly the case in situations, where people rely on these services for their basic needs. There is no evidence that a cyber blockade has ever resulted in death. However, should hospitals and emergency services go down as a direct result of the cyber blockade and result in loss of life, the cyber blockade may fall within the scope of the crime against humanity of "other inhumane acts". (Russell, 2014)

Additionally, cyber blockades that target critical infrastructure, in other words, those assets, systems, and networks that provide services necessary to carry on with daily life and even sustain life. For example, with the increasing digitalisation of power grids, they are becoming more and more vulnerable to a cyber blockade. (Krause et al, 2021) By blockading a power grid, perpetrators could block civilians from accessing necessary services, and even cause harm or death to individuals should a blackout occur at a hospital within its coverage.

For example, there has been evidence of a sophisticated attack on the University of Manchester in 2023, which involved a paralysis on NHS medical records and patient data. (Bateson, 2023) While this event cannot be classified as a cyber blockade, attacks like this within the scope of a cyber blockade could in turn bring the attack within the scope of Article 7(1)(k). Consequently, a cyber blockade targeting critical services or infrastructure could be characterised as "other inhumane acts".

4.2 Economic Impact

Cyber blockades that target financial institutions or disrupt the economy can cause a potential market crash, and even job loss with potential increase in poverty. This in turn could lead to significant suffering among the affected population. As technology currently stands, there appears to be no systemic cyber risk that would be capable of significantly impacting the economy, this is not proof of an absence of risk. (Warren, Kaivanto, and Prince, 2018) While Warren et al's article does not contemplate a cyber blockade, the piece rightfully concedes that their speculations are not risk free. Indeed, a cyber blockade could, theoretically, contribute to a significant economic downturn should its primary or sole aim be to paralyze the state economy. This, in turn, could have a detrimental effect on the respective civilian population to the extent that it may come within the scope of Article 7(1)(k).

This was the case, when the government of Costa Rica was compromised by a ransomware attack in 2022. Twenty-seven public institutions were targeted, which included the Ministries of Finance, Labour and Social Security, and the Science, Innovation, Technology and Telecommunications, the National Meteorological Institute, and the Social Security Fund. (Murry and Srivastava, 2022) The attack began with the crippling of the servers of the Finance Ministry, which disabled the Virtual Tax Administration and the Customs Information System. Then the websites of the Science Ministry and Meteorological Institution were disabled. The latter also saw information theft via an attack on the email server. A secondary attack, supposedly by a different actor, shut down the critical systems of the Social Security Fund, including the Single Digital Health Record. The only purpose was to extort money from the Costa Rican government. (Datta and Acton, 2023)

The government refused to pay the ransom, and as a result, the cyber blockade remained operational for several months until the end of June 2022. The government was forced to temporarily shut down the computer systems that are normally used for declaring taxes and for the control and management of imports and exports. This caused an economic loss of approximately US$ 125 million in the first two days after the cyber blockade was unleashed. (Chamber of Foreign Trade of Costa Rica, 2022) On 8 May 2022, the president of Costa Rica issued an executive order proclaiming a national emergency due to the cyberattacks against the country's public sector and stated that the country was in a "state of war". (Collier, 2022)

As the above example demonstrates, a cyber blockade could disable a state's financial sector and economy to a degree that it could be categorised as "other inhumane acts". Since this most likely would happen within the context of a widespread or systematic attack, the cyber blockade could be characterised as a crime against humanity.

4.3 Humanitarian Impact

Under international humanitarian legal rules, once an offer of aid by an impartial humanitarian organisation is accepted by interested parties, they must '…allow and facilitate rapid and unimpeded passage of humanitarian relief for civilians in need…' (Additional Protocol I, Article 70(2), which broadened previously adopted position in Geneva Convention IV, Article 23. See also ICRC Customary International Humanitarian Law Study 2005, Rule 55.) While there is academic support for the contention that blocking humanitarian aid may potentially constitute a crime against humanity, it is largely restricted to kinetic means, in other words, the physical interference or blocking of such aid. (Kraemer, Bhattacharya, and Dhrubajyoti, 2012)

In a cyber scenario, if the cyber blockade interferes with the delivery and administration of humanitarian aid, medicine, and other basic necessities to ensure survival, it can result in severe suffering and harm to those who depend on such assistance for their survival. (Tallinn Manual 2.0, 2017) The obligation not to interfere includes the obligation to ensure personnel are respected and protected: a cyber blockade must not be utilised to block the delivery of humanitarian aid and to cut off personnel from accessing aid to deliver. (Gisel, Rodenhauser, and Doermann, 2020) As such, blockading humanitarian assistance via cyber means could cross the threshold of Article 7(1)(k).

4.4 Psychological Impact

Prolonged cyber blockades, especially those that cut off communication or access to information, can lead to psychological distress, anxiety, and other forms of mental suffering among individuals who are isolated or unable to connect with loved ones.

Cyber blockades may trigger emotional trauma responses for there may be a sense of helplessness with respect to the extreme complexity of the cyberspace. (Kostyuk and Wayne, 2021) The amount of new technology put out on an almost daily basis makes it increasingly burdensome for lay civilians to keep up to date. Thus, this sense of rapid novelty potentially mixed with anxiety and dread may be the main driving force behind public responses. (McDermott, 2019)

Secondly, the ubiquity of cyberspace and digital interconnectedness, and the transnational reliance on the digital architecture mean that the civilian population may feel constantly threatened. Indeed, there are accounts of this from the 2007 Estonian Distributed Denial of Service attack. The editor of the Estonian daily newspaper Eesti Paevaleht, commented that 'You couldn't get information; you couldn't do your job. You couldn't reach the bank; you couldn't check the bus schedule anymore. It was just confusing and frightening.' (Mite, 2007)

Thirdly, the fact that there is increasing difficulty around deciphering who is behind a cyber blockade may cause additional negative psychological impact, which in turn can exacerbate the perceived risk. (Egloff 2020, Kaminska 2021) A sophisticated, high level cyber perpetrator would have little difficulty concealing their digital footprint, and could even make a cyber blockade look like as though it originated from somewhere else than in actuality. For example, should a cyber blockade have taken place, and the governmental and national security authorities are unable to trace the attack back to a perpetrator and make a confident attribution, this sense of uncertainty can exacerbate the psychological impact of the cyber blockade. As a result, given the potential emotional trauma, a cyber blockade may be characterised as "other inhumane acts" under Article 7(1)(k) of the Rome Statute.

5. Conclusion

Cyber blockades have evolved from the traditional law of blockades into the cyber sphere. Given a cyber blockade can be released in both war and peace, academia ought to examine its capabilities from a holistic perspective in the absence of a comprehensive regulatory framework. The most robust international legal framework that has a peacetime applicability is international criminal law, more specifically crimes against humanity. In contrast to war crimes, crimes against humanity can be committed during peacetime; unlike genocide, crimes against humanity is legally easier to establish for it does not require a special intent to kill based on group status.

However, as shown, the Rome Statute and its accompanying Elements of Crimes are becoming increasingly outdated for a cyber blockade would likely fall outside the scope of crimes against humanity. The only possible legal basis for international prosecution is Article 7(1)(k), which is designed to capture "other inhumane acts". Though, the International Criminal Court is yet to clarify its arguments for there is a discrepancy in jurisprudence from the ICTY and ICTR with reference to what is required to be established in order to successfully prosecute "other inhumane acts".

Thus, with the rapid emergence of new technologies that can be utilised in an offensive way whether in war or peace, international criminal law is increasingly becoming obsolete. A cyber blockade poses one of the greatest challenges to the existing framework of international criminal law given its destructive capabilities with respect to a civilian population and its respective state. Consequently, international criminal law must stay up-to-date with respect to unique cyber threats, such as cyber blockades. While theoretically these could be prosecuted internationally under Article 7(1)(k) of the Rome Statute, the International Criminal Court is yet to hear a case of this nature.

References

Associated Press, (2021) 'Colonial Pipeline confirms it paid $4.4m ransom to hacker gang after attack', The Guardian, 19 May.

Bateson S, (2023) 'NHS cyber attack follows rising numbers of data breaches in global healthcare sector' *Verdict*, 29 June.

Bhuiyan J, (2023) 'Cyberattack disrupts hospital computer systems across US, hindering services' *The Guardian*, 04 August.

Chan Yoon Onn K, (2023), 'The Prosecutor's New Policy on 'Cyber Operations' before the International Criminal Court (and its Implications for Ukraine): Some Preliminary Reflections' *EJIL Talk!* (15 September).

Chincilla D, (2022), '$125 millones en pérdidas estima Cámara de Comercio Exterior tras ciberataque que afectó aduanas' *Amelia Rueda*, 19 April.

Collier K, (2022) 'Costa Rica declares state of emergency over ransomware attack' *ABC News*, 11 May.

Datta, P. M., & Acton, T. (2023), "Ransomware and Costa Rica's National Emergency: A Defense Framework and Teaching Case" *Journal of Information Technology* Vol 0, No 0.

Drew P, (2017) *The Law of Maritime Blockade: Past, Present, and Future*, Oxford University Press.

Egloff F, (2020), "Public Attribution of Cyber Intrusions." *Journal of Cybersecurity* Vol 6 No 1.

Gisel L, Rodenhauser T, Doermann K, (2020) "Twenty years on: International humanitarian law and the protection of civilians against the effects of cyber operations during armed conflicts" (2020) *International Review of the Red Cross* Vol 102, No 913.

Henckaerts J-M and Doswald-Beck L, (2005) *Customary International Humanitarian Law Rules Vol 1*. Cambridge University Press.

Hyde C, (1922) *International Law Chiefly as Interpreted and Applied by the United States,* Little Brown.

Kaminska M, (2021), "Restraint under Conditions of Uncertainty: Why the United States Tolerates Cyberattacks." *Journal of Cybersecurity* Vol 7 No 1.

Kostyuk N, Wayne C, (2021), "The Microfoundations of State Cybersecurity: Cyber Risk Perceptions and the Mass Public." *Journal of Global Security Studies* Vol 6 No 2.

Kraemer J, Bhattacharya, Dhrubajyoti, (2012) "Blocking Humanitarian Assistance: A Crime Against Humanity?" *Leibnitz Information Centre for Economics*.

Krause T, Ernst R, Klaer B, Hacker I, Henze M, (2021) "Cybersecurity in Power Grids: Challenges and Opportunities" Special Issue Security and Trustworthiness in Industrial IoT Vol 21 No. 18.

Landler M and Markoff J, (2023) 'Digital Fears Emerge After Data Siege in Estonia' *The New York Times*, 29 May.

Marklund, Andreas, and Rüdiger, (2017) *Historicizing Infrastructure.* Aalborg University Press.

McDermott R, (2019), "Some Emotional Considerations in Cyber Conflict." *Journal of Cyber Policy* Vol 4, No (3), pp 309–25.

Mite V, (2007) Estonia: Attacks Seen As First Case Of 'Cyberwar' *Radio Free Europe*, 30 May.

Murry C, and Srivastava M, (2022) 'How Continue Ransomware Group Crippled Costa Rica - Then Fell Apart' *Financial Times*, 9 July.

Pomson O, (2023) ""Objects? 'The Legal Status of Computer Data under International Humanitarian Law", *Journal of Conflict and Security Law*, Vol 28, No 2, pp 349-387.

Renzo M, (2012) "Crimes Against Humanity and The Limits of International Criminal Law", *Law and Philosophy* Vol 31, No 4, pp 443-476.

Russel AL, (2014) *Cyber Blockades*, Georgetown University Press.

Schmidt M N et al, (2017) *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations*, Cambridge University Press.

Shandler R, Gross ML, Canetti D, (2023) "Cyberattacks, Psychological Distress, and Military Escalation: An Internal Meta-Analysis", *Journal of Global Security Studies* Vol 8 No 1.

Vojinovic I, (2023), '10 of the Biggest Data Breaches in History' *DataProt*, 14 July.
Warren P, Kaivanto K, Prince D, (2018) "Could a Cyber Attack Cause a Systemic Impact in the Financial Sector?" Bank of England Quarterly Bulletin Q4.
Legislation
Elements of Crimes 2011. Available at: https://www.icc-cpi.int/sites/default/files/Publications/Elements-of-Crimes.pdf (Accessed 18 October 2023).
Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I) 1977. Available at: https://ihl-databases.icrc.org/en/ihl-treaties/api-1977 (Accessed 18 October 2023).
Rome Statute of the International Criminal Court 1998, c. 2. Available at: https://www.icc-cpi.int/sites/default/files/RS-Eng.pdf (Accessed 18 October 2023).
Cases
Prosecutor v Akayesu (1998) ICTR-96-4-T.
Prosecutor v Bagilishema (2001) ICTR-95-1A-T.
Prosecutor v Bemba (2009)ICC-01/05-01/08.
Prosecutor v Bemba (2016) ICC-01/05-01/08
Prosecutor v Blaškić (2000) IT-95-14-T.
Prosecutor v Gbagbo and Blé Goudé (2012) ICC-02/11-01/15.
Prosecutor v Katanga (2012) ICC-01/04-01/07.
Prosecutor v Katanga and Ngudjolo Chui (2009) ICC-01/04-01/07 OA 8.
Prosecutor v Kayishema and Ruzindana (1999) ICTR-95-1- T.
Prosecutor v Kunarac et al (2004) IT-96-23-T & IT-96-23/1-T.
Prosecutor v Muthaura et al (2012) ICC-01/09-02/11.
Prosecutor v Tadić, (1997) IT-94-1-T.
Prosecutor v Stanišić (Jovica) and Simatović (Franko) (2017) MICT-15-96-PT, ICL 1099.

Covert Subversive Agents and Consensus Disruption on Large Projects

Johannes Vorster and Louise Leenen
University of the Western Cape, South Africa & Centre for Artificial Intelligence Research (CAIR), South Africa

jvorster@uwc.ac.za
lleenen@uwc.ac.za

Abstract: Is it possible to dramatically affect and influence military and other projects through social engineering of the consensus processes? In this paper we explore the impact that subversive agents can have on the ability of projects to move forward by disrupting the social cohesion and decision-making abilities of the processes designed to reach consensus. A consensus simulator is used to model group social cohesion behaviour in the context of project deliverables and show what the effect can be on the effort to reach consensus (number of meetings) as well as the time to reach consensus (calendar time) when subversive agents attempt to influence the groups making up the project team in such a way that it delays the ability of the team to reach consensus on key decisions. Many military options are available to delay enemy projects, including the assassination of enemy scientists, sanctions aimed at denying key project components, or even direct military action such as bombing the enemy facilities. However, this paper focusses on aspects of soft-force projection through covert disruption of project timelines. A social simulator was constructed that models individual agent's beliefs about various key topics within the context of a project. The effect that a small number of subversive agents can have on the time- and effort of a project is shown. In their covert actions, these subversive agents need to stay hidden, and thus their covert actions are limited, yet they can exert significant damage to the project in terms of delays. In this paper we present results showing the effects that such a small group can have, as well as pointing out that there seem to be a critical group size over which the subversive agents can not only have significant impact on project-delays but can also steer and direct certain key decisions.

Keywords: Subversive agents, Consensus, Consensus simulation, Multi-agent simulation, Stochastic consensus models

1. Introduction

Achieving consensus among diverse stakeholders is essential for mission success in large projects as well as military operations (Floyd, 1992; Charchian, 2001). However, this can be challenging in complex and dynamic environments where subversive agents (SAs) may intentionally disrupt the consensus-building process. For the purposes of this paper SAs are individuals or groups who promote views that cause more debate and delay the achievement of consensus and they dynamically change their views to purposefully delay the consensus process. It is important to distinguish between this subversive behaviour and normal healthy debate and disagreements, here the subversive agents promote views that they themselves may not believe, but they do so to create discord.

This paper explores the impact of SAs on consensus-seeking processes using a multi-agent simulator (MAS) and aims to provide insights into how organizations can mitigate their effects. The findings of this study have implications for Information Operations (IO) in military contexts, where achieving consensus among different stakeholders is critical for effective decision-making. The exploitation of project related social cohesion by subversive agents to affect project delivery has not been studied, to your knowledge.

1.1 Historic Perspective

IO involves the integration of various information-related capabilities to influence, disrupt, corrupt or usurp an adversary's decision-making process while protecting one's own. The presence of SAs can significantly impact IO efforts by delaying or disrupting consensus-building processes, leading to ineffective decision-making and potentially compromising mission success.

Therefore, understanding how SAs operate and their impact on consensus-building processes is critical for military planners and decision-makers involved in IO efforts. This paper builds on our previous research on consensus-seeking processes using MAS (Vorster & Leenen, 2023a, 2023b) and provides insights into how organizations can identify and mitigate the effects of SAs in IO contexts.

There have been several instances in military history where projects were delayed due to foreign influence, leading to inter-team conflict and discord. One such example is the development of the US F-35 Joint Strike Fighter program, which was delayed due to foreign influence from partner countries such as the United

Kingdom and Italy (Maldonado, 2015). These countries had specific requirements for the aircraft that were not initially included in the program, leading to disagreements and delays in the development process.

Another example is the development of the US M1 Abrams tank, which was delayed due to foreign influence from Germany, which had specific requirements for the tank's armour and weaponry (McDonald, 1995). This led to disagreements and delays in the development process, causing tension between the US and German teams working on the project. Other examples include the development of the Joint Strike Fighter (JSF)

These examples were not discord created by adversaries, but we can image that such conflict can be created by agents influenced by adversaries.

1.2 Healthy Conflict Within Project Teams

Earlier case studies of project failures have shown repeatedly that the top causes are project complexity and the failure of project team members to reach consensus on key points, be it requirements, design, or implementation aspects (Al-Ahmad et al., 2009; Whitney & Daniels, 2013; Kian, Sun, & Bosché, 2016; Waheeb & Andersen, 2022).

The process of reaching consensus implies that functionally diverse teams must understand the problem, surface, and share knowledge about the problem, identify any commonalities that may exist between views as well as differences in views, consolidate every difference in perspective, and through a team effort propose a solution that incorporates the final views of the whole team (Cheung et al, 2016). However, studies have also shown (eg Enyinda et al 2022) that healthy task conflict can increase a team's ability to reach consensus. This positive correlation between task conflict and team performance is especially notable in sufficiently complex tasks such as when there is no clear solution and involve significant uncertainty, such as during innovation and design, so that subjective factors determine the consensus outcome (Paletz et al 2017), neither do all problems have a single solution.

1.3 Subversive Agents Versus Healthy Conflict

Some remarks are needed on subversive agents versus conventional disagreement on topics between peers trying to find satisfactory solutions. In many projects the issues are complex, and people have justifiable reasons for arguing positions that others may consider unimportant. This is normal, and part of finding good solutions. The fact that there are debates imply that topics are not simple. Good communication, collaboration, and social skills within the milieu of task conflict has been found to be important for a team to reap the benefits from healthy conflict (Hirvonen, 2019). The rhetorical question "is there a way to tell the difference?", is an open question.

Debates within this context leads to longer times to reach consensus and to reject healthy conflict would lead to worse decisions being made. That is, the conflict creates points of debate that ultimately leads to better decisions. In this paper we use the term *subversive agent* to mean someone that is proactively trying to find arguments, trying to convince others of potions that they themselves do not believe in, specifically with the purpose of causing project delays and derailing decisions.

1.4 Social Influence Models

Early models of social influence were based on the effects that individuals with committed views were on the collective. A *committed view* refers to a view that is unchanging in the face of alternative views. Individuals with committed views have an unwavering belief in their opinion and they retain their view despite other individuals having other views. Over time, these individuals will convince others of their view, but since they are committed to their view, they will not change their position. So, the expected outcome is that eventually this committed view will take over as the dominant view.

Consider the following two example, the theory that the earth goes around the sun, as opposed to the over way around; and the spread of the conviction that female contraceptives should be used. There are many other examples the reader could identify, from the use of microwaves to the belief that smallpox vaccinations are good.

In both these cases the majority is convinced of something and then a new idea starts to propagate, driven by individuals with unwavering conviction.

We briefly mention two early influential models namely the *Bass (1969) diffusion model* and later *Granovetter's (1978) threshold model*. The diffusion model is based on product adoption, e.g. hairspray, where

the probability of an individual taking up the product is linear to previous buyers of the product. Adoption is driven by the number of other people that can be seen as having adopted. In the threshold model individuals observe the behaviour and views of their social neighbours, and if a certain threshold percentage is seen to have switched to the new view, then they will switch as well. The consequences of these models and their relevance to real-world problems has been studied extensively due to the importance of this topic to political and social views.

It was shown that there exists a critical threshold at about 10% of the population. If the number of individuals within the population that believe the new view exceeds 10%, then the time for the rest to be convinced is traumatically decreased (Xie et al, 2011). Later, Iacopini et al (2022) studied consensus processes through groups. In their model groups will gather and listen to a single speaker. They define a probability (beta, β) that the listeners will be convinced, as a group, and if that fails, everyone still has their own probability (alpha, α) that they may or may not be convinced. Let us call β the group's gullibility (our term, not theirs). They found that the ability of the committed minority to convince the entire population of their views falls into three distinct behavioural regions based on how gullible of the groups. Let us say that the initial majority view is represented by *view-a*, and the committed minority holds a view represented by *view-b*.

In the first region, with low β, the final equilibrium number of people with *view-b* is not the full population. If β is very small, then the equilibrium number of individuals that hold *view-b* is small. As β increases, the equilibrium number of individuals with *view-b* increases dramatically. At a specific critical β the equilibrium position reached is for the entire population to hold *view-b*. This ushers in region two where the minority can convince the entire population. As β is further increased, a second critical value is reached. For β values beyond this point the majority retain their *view-a*, and the minority is isolated to be the only individuals holding *view-b*.

1.5 Multi-Agent Systems in the Study of Consensus

We now turn to the more technical discussion on the use of Multi-Agent Systems (MAS). The past decade has seen significant studies on MAS and how to achieve and maintain consensus with a dynamically changing group of such agents under various conditions. For a detailed overview of this complex topic see reviews by Bao et al (2022) and Yang et al (2022). The applications of this research have been in high-speed consensus for Unmanned Aerial Vehicles (UAVs) and autonomous ground vehicles. One of the significant results is that if the difference of views between the various agents are measured, then that number decreases exponentially over time as the consensus protocol used drives agents to adopt the same views (Wei, 2021).

1.6 This Paper

In earlier papers we developed mathematical and simulation models based on stochastic processes in a MAS simulation to study consensus and subversive agent behaviour (Vorster and Leenen, 2023a; Vorster and Leenen 2023b). In this paper we aim to explain the behaviour of the models without using mathematics and stochastic terminology, but rather, focus on the implications of the results from the simulations and in particular the application within the IO and PsyOps environments, although some results can also apply to autonomous consensus, such as for self-driving cars and UAVs.

2. Characteristics of Consensus

In this section we discuss the general approach to modelling consensus using a multi-agent simulation system. First, we start with the consensus process without the presence of subversive agents. This is done to act as a baseline against which the behaviour of subversive agents can be juxtaposed.

We constructed a simulated multi-agent environment where agents must reach consensus on many topics and each agent starts the simulation with a stochastic distribution of views. Agents meet in pairs and discuss a random number of topics, from 1 to 10, with an average of 5.5 topics discussed per meeting. Agents will keep track of the topics and their view of each topic and understand if they differ in view on a topic from another agent. By comparing the view of agents pairwise with each other and summing the differences it becomes possible to construct a measure of the group consensus. Figure 1 (top graph) shows the trajectory of this consensus measure over time.

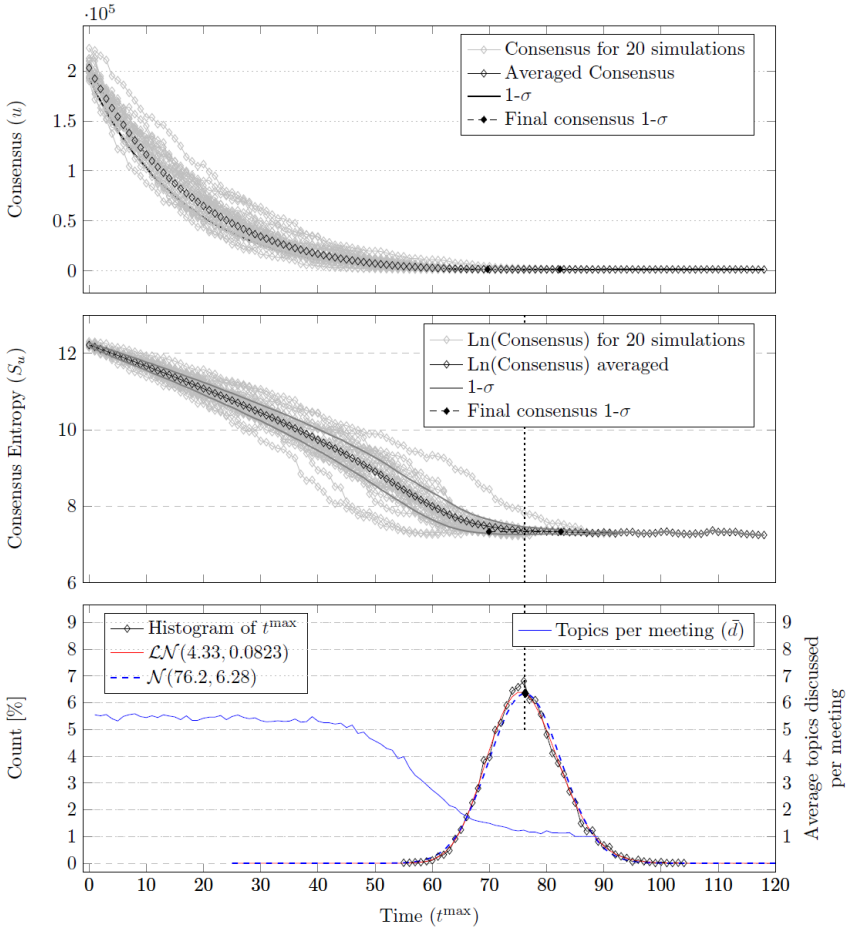


Figure 1: Measures of Consensus

Like earlier findings, the total consensus measure decreases exponentially (e.g. Wie, 2021). However, something that has emerged in our simulations that were not part of earlier consensus approaches is that real-world individuals need to meet and discuss topics, that is, spend time discussing matters. In the traditional MAS simulations, the agents communicate their views on all topics simultaneously to other agents. This is not realistic in a real-world scenario where it takes time to discuss topics and a given meeting has a limitation on the number of topics that can be discussed. The consequence of the limitations on number of topics that can be discussed can be seen when plotting this exponential decrease in consensus on a log-scale graph, Figure 1 (middle). If the decrease was exponential throughout the process, then the log-scale graph would remain linear, however, as can be seen in the graph, the consensus process is greatly affected by the number of topics available for discussion. Once two individuals have less than ten open topics for discussion, the meetings become much less efficient, and the consensus measure decreases more slowly.

The consequences of this finding for real-world consensus processes are that it may seem that the individuals are reaching consensus quickly on several topics, but the last few topics can take a significant time to resolve. Projects aiming to reach consensus on many topics involving many individuals my see early process in consensus measures but will quickly find a plateau of slow progress as many individuals must speak to many other individuals, but, only on a small number of topics. If one individual has 10 open items to discuss but with ten different individuals, it will take ten meetings, which is highly inefficient compared to when those ten items was related to only one or two other individuals and thus will need only a small number of meetings to resolve. The blue graph in Figure 1 (bottom) shows the evolution of the number of topics that are discussed per meeting. Initially the average is 5.5 = (10+1)/2 because a meeting will always discuss at least one topic and a maximum of ten topics. However, as time progresses, this number decreases and towards the end of the simulation meetings are dominated by one-topic discussions.

3. Subversive Agents

After a rather lengthy introduction, necessitated by the novelty of the topic, we now turn to subversive agents and their behavioural models.

Unlike the individuals with *committed views* discussed in the context of social influence models earlier, the aim of subversive agents is not to spread a specific view, but rather to spread views to increase the diversity of views, to promote views in a way that would lead to less consensus among individuals, thus delaying consensus formation on a group level. These agents may promote *view-a* to one individual and *view-b* to another individual, if that would lead to the two individual having views that evolve apart from each other.

Subversive agents also do not take extreme positions since that could lead to their exposure. They promote views already present within the group. That is, if it is known that some member holds *view-a*, then a subversive agent can promote that view to another member of the group. As the group's views evolve and contract, so do the positions promoted by the subversive agents.

3.1 Subversive Agent Behaviour

We quickly discuss the behaviour of subversive agents within the group. Since agents meet in pairs to discuss outstanding topics (topics where group consensus has not been reached) the subversive agents who understand the position of their counterpart can take up a different position and try to convince the counterpart of their view. Subversive agents select the view they choose to portray from the diversity of the group views in such a way that it falls outside the central views, but not so far way that they will seem to have radical views. Figure 4 below shows these three potential positions, and our subversive agents take up positions within the barred blue area illustrated in the figure. We investigated a number of distribution, but the results reported here is not dependent on distribution of views used, and so we only report on the findings using a Normal distribution to reduce complexity of reporting.

3.2 Small Numbers of Subversive Agents can Critically Harm Project Delivery

Now that the background context for subversive agents has been created, we turn to answer some questions about the effectiveness of such subversive agents. The first question that we want to answer is: can a single subversive agent have any real effect on project delivery?

Figure 2 shows the consensus trajectories for a group of twenty individuals among which are subversive agents. This is a graphical view of the effect that subversive agents have on the consensus process. As can be seen from the shift in slope of the graphs, subversive agents cause the consensus process to contract slower, and thus take much longer to complete.

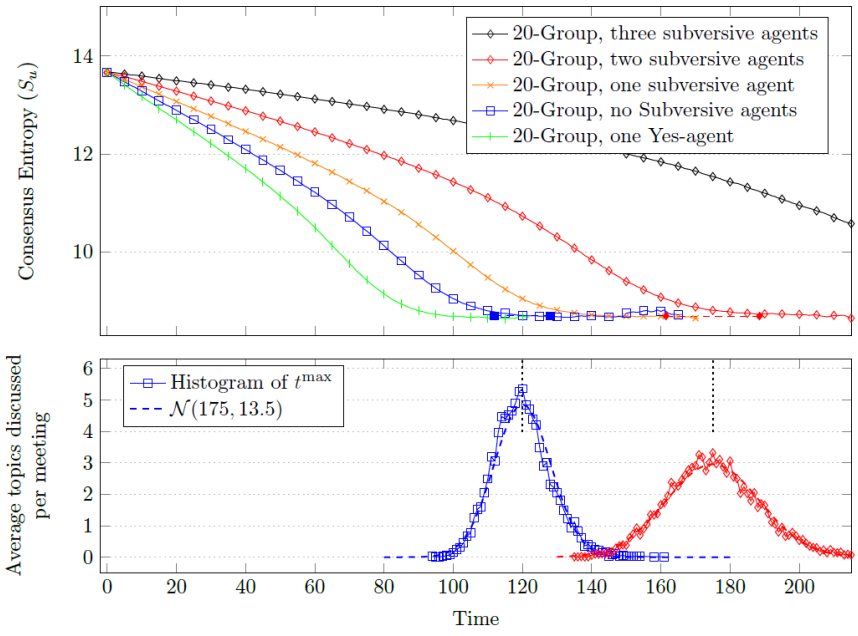


Figure 2: Impact of subversive agents on the time to reach consensus for a small number of contentious issues

If the project requires twenty individuals to complete, and it would have taken 120 time units to complete (hours, days, weeks), then the introduction of a single subversive agents will, on average, delay the project to take 139 time units (on average) to complete, which translates into a 16% project time delay, and thus also an increase of 16% in project costs. If two subversive agents are present, 10% of the group size, then they can cause a 48% increase in project costs and time, according to our simulations. Three subversive agents can more than double the project cost and time (127%), see Table 1.

Table 1: Subversive agents can extend the time to reach consensus significantly. This can be translated into project delays. The table shows the delays that subversive agents can cause on a project consisting of twenty individuals

Subversive Agents	Subversive Agents/Group %	Time to reach consensus	% Project delay (1σ)
0	0%	120	0% (14%)
1	5%	139	16% (14%)
2	10%	175	48% (16%)
3	15%	272	127% (22%)
4	20%	568	373% (42%)

3.3 Subversive Agent Cooperation

Since more than one subversive agent can be operationally active within the same project, one can ask if there are modes of cooperation between the agents that would improve their performance, as measured by their ability to delay the project?

Let us first consider ways in which these agents can cooperate. Firstly, (a) they can just not cooperate at all, that is, they act using their own information without being affected of by affecting other subversive agents. Intuition would dictate that this should be the least efficient way that agents can cooperate. There are at least two trivial other ways to cooperate, namely (b) for the agents to take up the same positions trying to sway more of the group to that position; and (c) for the agents to take up opposing views to attempt split the group into two factions.

Some of the results from simulations on these conditions are given in Table 2 below. There are too many datapoints to list, so we give a small representation in Table 2, and graph all the results in Figure 3.

The data from Table 2 suggests that all three these coordination types seem to have the same or close to the same effect. For two agents in a group of 20 using (a) no cooperation, (b) cooperation to take the same position, and (c) cooperation to take opposing positions, give delays of (a) 47.5%, (b) 46.2%, and (c) 49.0% respectively, each with a standard deviation of slightly over 15%. These values are so close that the maximum difference between then (3.8%) is well within the σ=15% boundary. These values are so close together that they look like a single point on the leftmost graph of Figure 3 (group-size 20 reading).

Table 2: Subversive agent coordination types and the effect that has on project delays

SAs/Group	Agent coordination	% delay (1σ)
2/20	No cooperation	47.5 (15.3)
	Same position	46.2 (15.6)
	Opposite position	49.0 (15.5)
4/40	No cooperation	75.3 (12.4)
	Same position	68.6 (12.1)
	Opposite position	70.4 (13.9)

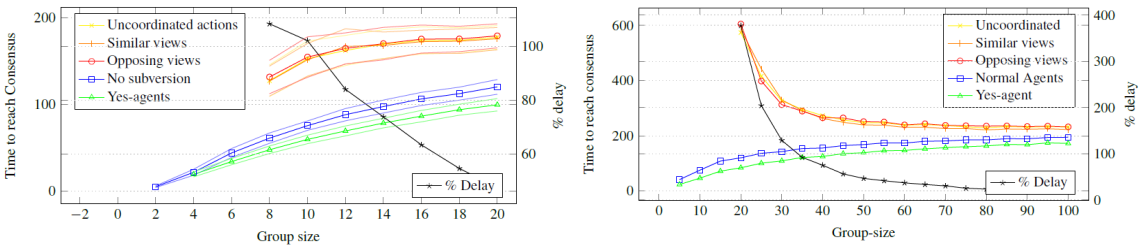


Figure 3: The graphs show the effect of cooperation between subversive agents. Three cooperative strategies are investigated, (a) uncoordinated agents, (b) same-view agents, (c) opposing-view agents. (Left) Two subversive agents operate within various group sizes. (Right) Four subversive agents operating

This is not an exceptional case and by visual inspection of Figure 3 (left) for group sizes 8 to 20 with two subversive agents and Figure 3 (right) for group sizes 20 to 100 with four subversive agents. This result is counter intuitive and warrants further analysis and discussion.

Let us assume that the views of all the individuals on a specific topic has a Normal distribution. Literature on this topic suggest that the distribution if highly dependent on the problem itself, and various distributions have been observed in the wild, form Normal, Uniform, Skew, Exponential, and even Double-normal (Den Boon and Van Meurs, 1991; Lang et al 2018). Our pic of distribution is illustrative and the argument we make will be valid for other distribution as well.

Any agent that takes a position within the centre 50% of the distribution of views, will thus promote a centrist view to the other 50% (see the central green area in Figure 4). If an agent takes up a view on the outside of this central region, that is the blue barred region or the outlying red region, then their efforts will increase the time to reach consensus. From the data we speculate that what is happening is that by taking positions in the outlying regions, the subversive agents widen the distribution, and it is this widening that is causing slower consensus formation. That is, it does not matter if agents expand the distribution of views to the left, or to the right, in both cases they are widening the distribution of views itself.

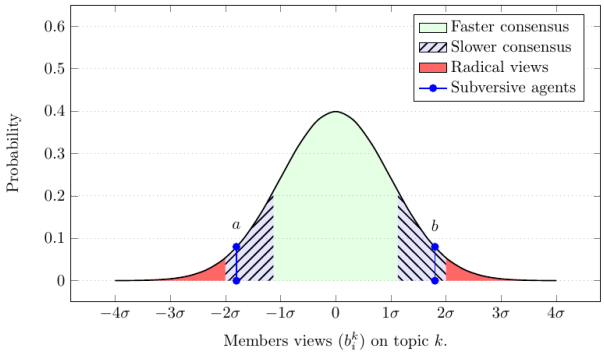


Figure 4: A theoretical distribution of views on a specific topic. If a subversive agent promotes a view in the central region (solid light green) then they are promoting consensus views, and it will lead to faster consolidation of views. If a subversive agent promotes a view

In Vorster and Leenen (2023b) we proposed a partial mathematical model to explain the results that agent cooperation does not affect the delay, but that model is not complete and thus some more work is needed to fully understand the observed behaviour. What the model and the data seem to imply is that subversive agents have the same effect no matter what their mode of working together is. It must be said that the irrelevance of cooperative model holds for the three modes we investigated. Maybe there are more sophisticated ways to cooperate that we did not investigate, that will yield better results.

For future work we propose the following additional cooperative models:

- Cooperate on specific topics, rather than on random topics. That is, maybe subversive agents will be more successful if they ignore topics with low diversity in views, and they should rather focus on specific topics with high diversity in views.

- Identify specific individuals within the group that have a high number of non-consolidated topics and focus to influence them specifically rather than targeting random members of the group.
- In our model topics are independent of each other, maybe if there were dependencies between topics subversive agents will find it easy to exploit one topic to create dependency-chaos in another topic.

4. Conclusion

In this paper we explored the potential effects that subversive agents can have on large projects such as is typically found in the military industrial complex.

Subversive agents have two goals, firstly, to extend the project life though delays in consensus formation within these large project teams, and secondly to not have such extreme views that they are seen as being difficult or sowing discord. They do this by promoting views that fall outside the central set of views, but too far outside. Subversive agents adapt their views as the group's views evolve, so that they eventually also agree to the consensus view.

We found that subversive agents can have a significant impact on the time to reach consensus in these projects, and a small number of subversive agents, 10% of the group, can easily delay a project by 45 to 50% for smaller groups, and by 70 to 75% for larger groups.

Furthermore, it seems from these simple models that the mode of cooperation between subversive agents do not have a significant impact on the delays they can achieve. This is an interesting result form a espionage perspective, because it implies that such agents or their handlers do not have to risk discovery through coordinated actions, uncoordinated actions will have the same effect.

References

Al-Ahmad, W., Al-Fagih, K., Khanfar, K., Alsamara, K., Abuleil, S., Abu-Salem, H. (2009). *A taxonomy of an it project failure: root causes*. International Management Review, 5 (1), pp 93-116.

Bao, G., Ma, L., Yi, X. (2022). *Recent advances on cooperative control of heterogeneous multi-agent systems subject to constraints: A survey*. Systems Science & Control Engineering, 10 (1), pp 539–551.

Bass, F.M. (1969). *A new product growth for model consumer durables*. Management science, 15 (5), pp 215–227.

Charchian, Daniel J. (2001). *Understanding Culture and Consensus Building: Requisite Competencies for Interagency Operations*. US Army War College, Carlisle Barracks, PA, USA. Online: https://apps.dtic.mil/sti/pdfs/ADA390805.pdf, accessed 2024/01/22.

Cheung, S.Y., Gong, Y., Wang, M., Zhou, L., Shi, J. (2016). *When and how does functional diversity influence team innovation? the mediating role of knowledge sharing and the moderation role of affect-based trust in a team*. Human relations, 69 (7), pp 1507–1531.

Den Boon, A. K. and Van Meurs, A. (1991). *Measuring opinion distributions: An instrument for the measurement of perceived opinion distributions*. Quality and Quantity, 25(4), pp 359–379.

Enyinda, C.I., Blankson, C., Cao, G., Enyinda, I.E. (2022). *Why cannot we all just get along? resolving customer-focused team interface conflicts in a b2b firm leveraging ahp-based multi-criteria decision-making*. Journal of Business & Industrial Marketing, , https://doi.org/10.1108/jbim-02-2021-0104.

Granovetter, M. (1978). *Threshold models of collective behavior*. American journal of sociology, 83 (6), 1420–1443.

Floyd, Steven W., and Bill Wooldridge. (1992). *Managing strategic consensus: the foundation of effective implementation*. Academy of Management Perspectives, 6 (4), pp 27-39.

Hirvonen, P. (2019). *Positioning, conflict, and dialogue in management teams*. Qualitative Research in Organizations and Management: An International Journal, 14 (4), pp 444–464.

Kian, M.E., Sun, M., Bosché, F. (2016). *A consistency-checking consensus-building method to assess complexity of energy megaprojects*. Procedia-social and behavioral sciences, 226 , pp 43–50.

Lang, J. W., Bliese, P. D., and de Voogt, A. (2018). *Modeling consensus emergence in groups using longitudinal multilevel methods*. Personnel Psychology, 71(2), pp 255–281

Maldonado, M. M. (2015). *Qualitative Case Study on F-35 Fighter Production Delays Affecting National Security Guidance*. Doctoral dissertation, Walden University.

McDonald, B. N. (1995). *Evaluating foreign-source dependencies in the US Army's M1 Abrams tank*. Doctoral dissertation, Monterey, California. Naval Postgraduate School.

Paletz, S.B., Chan, J., Schunn, C.D. (2017). *The dynamics of micro-conflicts and uncertainty in successful and unsuccessful design teams*. Design Studies, 50 , pp 39–69.

Vorster, J., & Leenen, L. (2023a). *Consensus simulator for organisational structures*. Proceedings of the 13th international conference on simulation and modelling methodologies, technologies, and applications (pp 15–26).

Vorster, J., & Leenen, L. (2023b). *Exploring the effects of subversive agents on consensus-seeking processes using a multi-agent simulator*. Proceedings of the 13th international conference on simulation and modelling methodologies, technologies, and applications (pp 104–114).

Waheeb, R.A., & Andersen, B.S. (2022). *Causes of problems in post-disaster emergency re-construction projects—Iraq as a case study*. Public Works Management & Policy, 27 (1), pp 61–97.

Wei, Qinglai, Xin Wang, Xiangnan Zhong, & Naiqi Wu (2021). *Consensus control of leaderfollowing multi-agent systems in directed topology with heterogeneous disturbances*. In: IEEE/CAA Journal of Automatica Sinica 8.2, pp 423–43.

Whitney, K.M., & Daniels, C.B. (2013). *The root cause of failure in complex it projects: Complexity itself*. Procedia Computer Science, vol 20 , pp 325–330.

Xie, J., Sreenivasan, S., Korniss, G., Zhang, W., Lim, C., Szymanski, B.K. (2011). Social consensus through the influence of committed minorities. Physical Review E, 84 (1), 011130.

Yang, R., Liu, L., Feng, G. (2022). *An overview of recent advances in distributed coordination of multi-agent systems*. Unmanned Systems, 10 (03), pp 307–325.

Exploring South Africa's Cybersecurity Legal Framework Regulating Information Confidentiality, Integrity, and Availability

Murdoch Watney
University of Johannesburg, South Africa

mwatney@uj.ac.za

Abstract: Globally the digital ecosystems of all countries face a common denominator, namely the threat of cyber operations and how to respond effectively to the threat. There are various types of cyber operations, but the discussion focusses on the national cybersecurity legal framework regulating non-state criminal cyber operations that target the confidentiality, integrity, and availability of information. It does not deal with offensive state or state-sponsored cyber operations which fall within the remit of the international law. As Internet access continues to grow and daily life becomes increasingly dependent on Information and Communication Technologies (ICTs), users, institutions, businesses and government are exposed to a variety of cybersecurity threats. The cyberthreats – with specific reference to threats to the confidentiality, integrity and availability of information (data) - are real. For example, one single security breach may result in the exposure of the personal information of millions of people and that exposed information may be used to commit other crimes, such as theft, fraud or extortion. The first line of defense to a criminal non-state cyber operation targeting the confidentiality, availability and integrity of information is a robust and resilient cybersecurity strategy. This strategy must provide amongst others for a cybersecurity legal framework that provides for the: identification of different types of cyber threats to the confidentiality, integrity, and availability of information; and legal response to threats that result in a security compromise (data breach) which also constitutes a cybercrime. The cybersecurity legal framework must provide for the following interlinked key concepts, namely cybersecurity, privacy, data protection, data breach and cybercrime. The discussion will highlight the meaning and relevance of the concepts and the impact of artificial intelligence (AI). It is against above background that the South African legal cybersecurity framework is explored to establish whether it effectively regulates cyber threats targeting information confidentiality, integrity, and availability. Evaluating cybersecurity incidents and the legal response to it may also contribute to a better understanding of the global challenge cyberthreats present to confidentiality, integrity, and availability.

Keywords: Cybersecurity, Privacy, Data protection, Cybercrime, Cyberthreat, Cybersecurity legal framework

1. Introduction

Various entities, such as government, military, corporate, financial, medical facilities and education institutions collect, process, and store unprecedented amounts of data on computers and other devices. A significant portion of that data may be sensitive information, for which unauthorized access or exposure could have harmful and legal consequences (De Groot, 2023). A security compromise impacts negatively on the reputation and trust of the entity, recovering from the data breach has financial implications and the entity may also face legal consequences. In 2023 the average data breach cost for South African organisations reached R49.45 million (Tredger,2023).

As the volume and sophistication of criminal non-state cyberoperations continue to grow, companies and organizations, especially those that are tasked with safeguarding information relating to national security, health, or financial records, have to take steps to protect personal information (De Groot, 2023; Patterson, 2023). A cybersecurity legal framework that provides for cybersecurity, privacy, data protection, cyberthreat and cybercrime is therefore essential in ensuring the protection of information against non-state criminal cyber operations that specifically target the confidentiality, integrity and availability of information.

For purposes of this discussion, the following South African cybersecurity legislation will be deliberated against the background of the Constitution:

- The Protection of Personal Information Act (POPIA) 4 of 2013; and
- The Cybercrimes Act 119 of 2020.

The discussion explores the effectiveness of the South African cybersecurity legal framework in responding to the threat and risk of a non-state criminal cyber operation that targets the confidentiality, integrity, and availability of information.

2. Understanding the Relationship Between Cybersecurity, Privacy, Data Protection, Cyberthreat, Cybercrime and Impact of Artificial Intelligence (AI)

A government must provide a secure and trustworthy digital ecosystem aimed at protecting information, otherwise a society cannot prosper, nor can the economy thrive. In protecting the digital ecosystem, cybersecurity would be the first line of defence.

Cyber security is an umbrella term that refers to a collection of technical tools, processes, policies, laws, training and awareness programmes etc. that are used to protect systems such as hardware, software and information connected to the internet from threats in cyberspace (Patterson, 2023). The discussion focusses specifically on the legal protection of the confidentiality, integrity and availability of information against cyberthreats.

Section 1 of the South African National Cybersecurity Framework Policy (NCPF) of 2015 defines cybersecurity as:

- The practices of making the networks that constitute cyberspace secure against intrusions;
- Maintaining confidentiality, integrity and availability of information;
- Detecting intrusions and incidents:
- Responding to; and
- Recovering from them.

The concepts, cybersecurity, privacy, data protection, cyberthreat and cybercrime are interlinked. For purposes of this discussion, the term, "intrusion", is replaced with "cyberthreat". Cybersecurity is aimed at protecting privacy and data against cyberthreats, such as unauthorised access to personal information. A cyberthreat may manifest itself in a cybercrime. Cybersecurity consists of pro-active measures aimed at preventing a cyberthreat. Cybersecurity is, however, not infallible and if a security compromise manifests itself as a crime, cybercrime legislation - which is reactive by nature - must provide for the criminalisation of such conduct. A cybercrime may be defined as a criminal offence that is facilitated by or involves ICTs (Watney, 2022).

In terms of South African law, the right to privacy is protected in terms of the common law and section 14 of the Constitution of South Africa 1996. Section 14(d) provides for the non-violation of the privacy of communications. POPIA recognises the right to privacy enshrined in the Constitution and gives effect to this right through mandatory procedures and mechanisms for the collection, use, and processing of personal information. POPIA is essentially, the South African Data Protection Act.

The impact of artificial intelligence (AI) systems in respect of cybersecurity, data protection and cybercrime must be considered. AI systems have advanced and often rely on vast data to train their algorithms and improve performance. This data can include personal information such as names, addresses, financial information, and sensitive information, such as medical records. The collection and processing of this data can raise concerns about how it is being used and who has access to it (ETsystems, 2023). Organisations that deploy security AI and automation extensively experience, on average, nearly R10.49 million lower data breach costs than organisations that do not deploy these technologies (Tredger, 2023). On the other hand, threat actors are also making use of AI (Legodi, 2023).

3. South Africa's Legal Cybersecurity Framework Regulating Information Confidentiality, Integrity and Availability

3.1 Background to the Legal Cybersecurity Framework

Today's digital ecosystem evolves around information. Information has not only become one of the world's most valuable intangible assets, but it is also globally one of the most vulnerable assets (Eldridge et al, 2023). If someone gains unauthorised access to personal information, such information can be used to commit a number of crimes, for example theft, fraud and extortion. It is important that the responsible person has a legal duty to protect personal information and if the responsible party does not comply with the legal duty, it should be held liable. Furthermore, the conduct that compromises the confidentiality, integrity and availability of information as well as the conduct that uses that information to commit other crimes, must be criminalised to ensure the threat actor is held criminally liable.

Prior to the Protection of Information Act 4 of 2013 (POPIA), personal information was inadequately protected in South Africa. South Africa was in dire need to implement data protection legislation to give effect to the constitutional right to privacy. In July 2021, data protection legislation, POPIA, came into effect. It brough South

Africa in line with current international trends and laws on privacy (Malinga, 2023). POPIA is similar to the European Union (EU) General Data Protection Regulation (GDPR).

It does not help to provide for the protection of personal information in accordance with POPIA if the conduct relating to the confidentiality, integrity and availability of information and conduct that uses the unlawfully obtained information to commit crimes are not criminalised. The Cybercrimes Act 19 of 2020 (Cybercrimes Act) which came into effect in December 2021 criminalises not only conduct that contravenes confidentiality, integrity and availability but also conduct relating to the use of such information to commit other offences. For example, the South African pharmaceutical company, Dis-Chem indicated after it had suffered a security breach in 2023, that there was a possibility that the threat actor could use the impacted personal information for further criminal activities such as phishing attacks and e-mail scams (Illidge, 2023). In August 2023, the online newspaper, Daily Maverick suffered a Distributed Denial-of-Service attack (DDoS). It received more than 36 million hits which emanated from a single domain (O'Regan, 2023).

The theoretical interaction between POPIA and the Cybercrimes Act and the relevant statutory sections regarding information confidentiality, integrity and availability will be discussed hereafter at par. 3.2 and par. 3.3 respectively. At par. 4 the practical application of the relevant legislation is deliberated by exploring the legal response to security compromises and whether there are legal shortcomings that necessitate attention.

3.2 Protection of Information Act 4 of 2013 (POPIA)

POPIA (available at https://popia.co.za) and the Regulations relating to POPIA (POPIA Regulations; see https://www.popiact-compliance.co.za/images/Documents/POPIA Regulations - Dec 2018.pdff) are South Africa's primary data protection laws. POPIA provides that a responsible party such as any organisation or person who processes personal information of a data subject must take security measures to prevent unauthorised access, loss, damage, and destruction of personal information. POPIA further provides for the establishment of a data protection authority, the Information Regulator, who is an independent supervisory authority which has been established for the purpose of administering POPIA.

"Personal information" and "processing" are defined in section 1 of POPIA. "Personal information" is "any information relating to an identifiable, living natural person; or an identifiable, existing juristic person" such as education, physical health, mental health, finances, criminal or employment history; ID number, symbol, email address, physical address, telephone number, location or online identifier; names, surname, birth date or beliefs, and biometrics. POPIA has a wide application and impacts all persons processing personal information (Mokdad et al, 2023). "Processing" refers to the "collection, recording, organisation, storage, updating or modification, retrieval, consultation, use, dissemination by means of transmission, distribution or making available in any other form, merging, linking, as well as blocking, erasure or destruction of personal information".

POPIA provides eight information protection conditions (principles) that govern the processing of personal information. In terms of condition 7, a responsible party has a legal duty to protect personal information against a security breach, also referred to as data breach or security breach. POPIA does not define a data breach but it constitutes the unauthorised access and disclosure of personal information to a third party without approval. Although condition 7 of POPIA does not stipulate the exact measures that must be taken, it requires that the responsible party must take appropriate and reasonable cybersecurity measures to prevent a loss or compromise of personal information.

Section 22 of POPIA states that where there are reasonable grounds to believe that the personal information of a data subject has been accessed or acquired by an unauthorised person, the responsible party must report the breach to the Information Regulator, and to the data subject. Unlike other privacy laws around the world, such as the GDPR, there is no "harm" or "impact" threshold to trigger a responsible party's reporting obligations under POPIA (Boda et al, 2023). Even if a compromise is unlikely to cause harm or loss to a data subject, reporting obligations still apply. This means that seemingly low-risk events, for example a stolen work laptop but which is subsequently locked by the company's IT department to prevent unauthorised access, or an email containing personal information sent to the wrong person but successfully recalled before being read, fall within the scope of reportable security compromises (Boda et al, 2023). In August 2022, the Information Regulator published a notification template and guidance to facilitate the reporting of security compromises in terms of section 22 of POPIA (see https://inforegulator.org.za/wp-content/uploads/2020/07/Guidelines-on-completing-a-Security-Compromise-Notification-ito-Section-22-POPIA.pdf).

The Information Regulator will investigate if the responsible party had taken reasonable steps to protect the information. If a responsible party is alleged to have committed an offence in terms of POPIA, the Regulator may

issue the responsible person (referred to as the infringer) an infringement notice with instructions to rectify the security compromise. POPIA sets down firm frameworks that organisations have to abide by to avoid fines, criminal prosecution and potential reputation loss. Perpetrators who do not adhere to the enforcement notice can face fines of up to R10 million or 10 years imprisonment, depending on the seriousness of the breach.

Non-compliance with the provisions of POPIA or an enforcement notice may therefore result in the Information Regulator holding the responsible person civilly and criminally liable for a security breach and non-compliance with the enforcement notice. The individuals responsible for the security breach can also face criminal prosecution and conviction and be sentenced to imprisonment for a period not exceeding 10 years (Boda et al, 2023). The data subject can also institute a civil claim against the responsible party for non-compliance with

3.3 Cybercrimes Act 19 of 2020 (Referred to as Cybercrimes Act)

Conduct that affects the confidentiality, integrity and availability of information, such as unauthorised access to information, ransomware attacks, brute force attack and business email compromise (BEC) must be criminalised.

Chapter 2 part 1 sections 2 – 7 of the Cybercrimes Act (available at https://www.gov.za/sites/default/files/gcis_document/202106/44651gon324.pdf) provides for crimes against the confidentiality, integrity and availability of information, such as unauthorised access to information; unlawful interception of data, unlawful acts in respect of software or hardware to gain unauthorised access to information or intercept data or unlawfully interfere with data or a computer program or computer data storage or computer system or to unlawfully acquire or use a password to access information. The information obtained in such an unlawful manner can be used to commit other crimes such as those provided for in part 2 sections 8 – 10 and 12 of the Cybercrimes Act, namely fraud, forgery and uttering, extortion and theft.

4. Legal response to a Data Breach and Cybercrime

4.1 Examples of Legal Response

The South African theoretical legislative cybersecurity framework is commendable, but the effectiveness of the legal framework can only be evaluated by means of the legal responses to security compromises.

Since the implementation of POPIA and the Cybercrimes Act, there have been various security breaches. The legal response to the data breach by the responsible party and information Regulator in terms of POPIA and the response by the police to the cybercrime in accordance with the Cybercrimes Act will be explored with reference to various examples.

Example 1: The Information Regulator investigated security compromises at two credit bureaus, namely Experian and TransUnion, which experienced data breaches in 2020 and 2022 respectively (Labuschagne, 2023). TransUnion and Experian are major credit bureaus that hold vast amounts of personal data in their databases. This can include names and surnames, identification numbers, passport numbers, contact information, home addresses, credit histories, dates of birth, vehicle finance contract numbers, and car VINs. Companies and banks use the data to conduct credit vetting on potential customers, as well as other financial background checks.

The Experian data breach occurred in 2020 prior to the implementation of POPIA. It was first reported by the South African Banking Risk Centre (SABRIC) (Labuschagne, 2023). Experian detected the breach more than 50 days after the data had already been transferred (Molosankwe, 2023). The incident exposed as many as 24 million South Africans and nearly 794,000 business entities which dataset the threat actor, Karabo Phungula obtained under false pretences. Phungula allegedly wanted to sell the data for R4 million. He was arrested in 2021, pleaded guilty to fraud and was duly convicted of fraud in 2022 and sentenced to 15 years imprisonment in 2023 (Molosankwe, 2023).

In 2022 TransUnion experienced a ransomware attack. A ransomware attack encrypts data, rendering it inaccessible unless a specified ransom is paid to the threat actor in exchange for decryption (Mostert, 2023). Ransomware attacks pose a serious threat globally, for example in 2021 the United States of America (US) experienced the Colonial Pipeline ransomware attack which severely disrupted its supply chain (Watney, 2022). In 2022 approximately 78% of South African organisations experienced ransomware attacks (Mostert, 2023).

In the TransUnion security breach, the threat actors reportedly obtained personal information of approximately 5 million individuals. A Brazilian hacking group named N4ughtySec claimed responsibility for the security compromise. TransUnion allegedly refused to pay a $15-million (R224 million at the time) ransom to prevent the data being leaked online (Labuschagne, 2023).

Example 2: In September 2021, the Department of Justice and Constitutional Development (DoJ and CD) suffered a ransomware attack on its IT systems, leading to all its information systems being encrypted and unavailable to internal employees, as well as members of the public. As a result of the attack, all electronic services provided by the Department were affected, including the issuing of letters of authority, bail services, e-mail and the departmental website. During the ransomware attack at least 1 200 files containing the names, banking details and contact details of those who had submitted personal information were compromised (Illedge, 2023; Vermeulen, 2023).

On 9 May 2023, the Information Regulator issued an infringement order in which it found that various sections of POPIA had been contravened. The Information Regulator found that the Department had been negligent as it failed to put in place adequate technical measures to monitor and detect unauthorised exfiltration of data from their environment resulting in the loss of approximately 1 204 files. The enforcement notice required the Department to submit proof to the regulator within 31 days of receipt of the notice that the Trend Anti-Virus licence, the SIEM licence and the intrusion detection system licence had been renewed. It also required the Department to institute disciplinary proceedings against the officials who had failed to renew the licences, which were necessary to safeguard the department against security compromises. An enforcement notice is legally binding but in terms of section 97(1) of POPIA, a responsible party may appeal against the enforcement notice. The Department did not appeal against the enforcement notice nor did it comply with the notice within the prescribed time period. As a result of the non-compliance with the enforcement notice, the Information Regulator issued an infringement notice to the Department in July 2023, in which it ordered it to pay an administrative fine of R5 million following its failure to comply with the enforcement notice issued by the regulator on 9 May 2023 (Illidge, 2023; Moyo, 2023; Vermeulen, 2023). The Department became the first institution to be fined for failure to comply with an enforcement notice. In October 2023, the Department indicated that it was challenging both the infringement and enforcement notices (Ndlovu, 2023).

Example 3: In October 2022, eight women were gang raped at an abandoned mine site in Krugersdorp. In order to apprehend the perpetrators, personal information of the victims such as their surnames ages, physical addresses were shared on the police WhatsApp group. Unfortunately, this information was unlawfully disclosed on social media and became public (Goba, 2023). Not only did this unauthorised disclosure cause the victims and family a lot of stress, but it may deter rape victims to report such crime out of fear their information will not be kept confidential. The Information Regulator found that the police had violated POPIA by not taking reasonable security measures to prevent the personal information from being disclosed to the public. In April 2023, the Information Regulator issued an enforcement notice ordering the responsible party to notify the data subjects of the security compromise which relates to their personal information, make a public apology and to investigate how this information was leaked to the public. In May 2023 the police commissioner issued a public apology to the rape victims (see https://www.gov.za/speeches/saps-apologises-krugersdorp-rape-victims-release-their-personal-information%C2%A0-3-may-2023).

Example 4: In May 2023, a well-known South African pharmaceutical company, Dis-Chem experienced a brute force attack (Illidge, 2023). A brute force attack is aimed at cracking a password by continuously trying different combinations until the right character combination is found (Illidge, 2023). On 1 May 2023, Dis-Chem became aware that 3,7 million of its clients' records were compromised when employees received SMS of the breach. It reported the breach to the Information Regulator on 5 May 2023. The Information Regulator issued Dis-Chem with an enforcement notice on 31 August 2023, instructing it to address its security issues or face a R10 million fine. The Information Regulator determined that the responsible party, Dis-Chem had failed to notify its data subjects as required in terms of section 22 of POPIA which the responsible party strongly disputed. It indicated that a formal notice of the security breach had been published on the Dis-Chem website and a media statement had been released nationally. (Mzekandaba, 2023; see https://inforegulator.org.za/wp-content/uploads/2020/07/FINAL-MEDIA-STATEMENT-ENFORCEMENT-NOTICE-ISSUED-TO-DISCHEM-PHARMACIES-LTD.pdf).

4.2 Lessons Learnt From the Legal Response

Globally unauthorised access to personal information and the use of that information for the commission of various types of cybercrime pose a serious challenge to all entities. The examples discussed at par. 4.2 illustrate the type of threats entities face, how they are committed as well as trends or changes in respect of the threats. Furthermore, by evaluating the legal response to security breaches, shortcomings on the side of the responsible party or police may be identified which can be rectified going forward.

South Africa has experienced an increase in cyberattacks on government entities, for example, both the Department of Defence and the State Security Agency (SSA) experienced security compromises in August 2023 The government has a mandate to safeguards its constituents against threats (Mostert, 2023). In this regard, the ransomware attack on the Department of Justice and Constitutional Development (Department) raises serious concerns regarding the manner in which the critical infrastructure was protected and the government commitment to securing the personal information of its citizens. The security breach was the result of negligence of the side of the Department who had not renewed the relevant security software. The Department did not comply with the infringement notice issued by the Information Regulator with the consequence that the Department was fined. Such conduct does not enhance trust that the government is taking its obligation to protect information seriously. Trust can only be ensured through robust internal cybersecurity measures, such as a cybersecurity policy that outlines the entities' data encryption, access controls, incident response, and preventive measures.

Mzekandaba (2023) refers to Jason Jordaan, a principal forensic analyst at DFIR Labs, who has expressed concerns in respect of the enforcement of cybersecurity legislation. Jordaan raises the following 2 points of concern:

- The police who are tasked with the investigation of a cybercrime, such as identifying the threat actor (perpetrator), may not have the relevant investigative experience to effectively investigate a matter, especially in circumstances where the security compromise is committed outside South Africa. For example, the 2022 TransUnion ransomware attack was committed outside South Africa and such cross-border investigation is very challenging despite the Cybercrimes Act read with the Criminal Procedure Act 51 of 1977 providing for the investigation. In many instances, the threat actor is not prosecuted. It is concerning that only an estimated 10% of cybercrime is reported (Kahle, 2023).
- Since the implementation of POPIA, many security compromises have been reported to the Information Regulator who may need assistance in the investigation in establishing whether the responsible party complied with the provisions of POPIA. The Regulator has a dual function and in addition to being the POPIA watchdog, the Regulator also oversees the enforcement of the Promotion of Access to Information Act, 2000 (PAIA) (Boda et al, 2023). South African law requires every organisation to have a PAIA manual available on their website which ensure legal compliance and enhances accountability.

Although above-mentioned concerns are valid, it is not unique to South Africa. The sophistication and number of cyber threats continue to increase which pose a serious challenge to all parties that process personal information on global level (Legodi, 2023).

The 2020 Experian security compromise illustrates that early detection and fast response can significantly reduce the impact of a breach (Tredger, 2023). The Dis-Chem security breach highlights that a responsible party who outsources processing to a third-party provider, will not escape liability. This was also the case when Britain's Financial Conduct Committee (FCC) fined Equifax in October 2023 for a data breach suffered in 2017. The United Kingdom (UK) arm of Equifax had outsourced personal information of British customers to the parent company, Equifax Inc, in the United Nations. FCC found that the UK arm of Equifax had not taken the necessary steps to protect British customers' data (Reuters, 2023).

5. Recommendations

The following is recommended:

- It is commendable that the Information Regulator issued reporting guidelines in 2022, but it should provide a specific time period in which a security compromise must be reported to the Information Regulator.

Article 33 of the GDPR specifies that an organization must report a security breach that affects personal data to a Data Protection Authority (DPA) within 72 hours of becoming aware of the breach (see https://edps.europa.eu/data-protection/our-role-supervisor/personal-data-breach_en). Likewise part 3 of the United Kingdom Data Protection Act of 2018 also provides that organisations must report certain types of personal data breaches within 72 hours of becoming aware of the breach to the Information Commissioner (https://ico.org.uk/for-organisations/law-enforcement/guide-to-le-processing/personal-data-breaches/). The US does not have general federal legislation impacting data protection, but there are a number of federal data

protection laws that are sector-specific which provides for mandatory reporting (see https://iclg.com/practice-areas/data-protection-laws-and-regulations/usa).

- There should be a guideline in respect of paying ransomware. Paying of ransomware should be discouraged. If a company pays ransom, it must be reported to either the Information Regulator or police in order to keep record of payments made.
- The reasons for the lack of reporting cybercrimes must be investigated. Can it be ascribed to the public's mistrust in the police having proper training and investigative experience to deal effectively with cybercrimes (Kahle, 2023)?
- Laws must keep up with technological advancements, such as the use of AI, generative AI and machine learning. South Africa does not have AI specific legislation. Current cybersecurity legislation should give effect to these advancements . POPIA has adopted a "privacy by design" principle which means it embeds privacy into the design. The resulting consequence in dealing with privacy issues is a system that is proactive and preventative, as opposed to reactive or remedial. This design must also take into account technological advancements, such as AI in respect of processing personal information (De Wet and Fourie, 2023; Malinga, 2023). The Information Regulator could issue a code of conduct in respect of the use of AI for personal information processing (de Wet and Fourie, 2023).

6. Conclusion

South Africa's theoretical cybersecurity legal framework provides for the protection of personal information as well as criminal conduct relating to the confidentiality, integrity and privacy of information. Implementation and enforcement of legislation such as POPIA and the Cybercrimes Act may be challenging. Both legislations are relatively new and consequently much can be learnt from exploring the legal responses to cybersecurity compromises. The Information Regulator has made inroads since POPIA came into operation and has taken decisive steps to issue infringement notices to the responsible parties that did not comply with the provision of POPIA. However, the lack of reporting and prosecution of cybercrimes will have to be addressed (Kahle, 2023).

A major challenge facing cybersecurity and privacy laws – not only in South Africa, but globally – is technological advancements and for purposes of this discussion, the legislature has to be cognisant of the impact of AI on cybersecurity, privacy, data protection and cybercrime and to ensure current legislation can adjust to the changes or implement new legislation if necessary.

References

Boda, R et al. (2023) "POPIA enters the terrible twos", [online], https://www.ensafrica.com/news/detail/7266/popia-enters-the-terrible-twos-.

De Groot, J. (2023) "What is cybersecurity? Definitions best practices and examples", [online], https://www.digitalguardian.com/blog/what-cyber-security.

De Wet, PR. and Fourie, J. (2023) "South Africa: AI and data privacy regulations – the complexities of AI technologies and processing personal information", [online], https://vdt.co.za/popia/south-africa-ai-and-data-privacy-regulations-the-complexities-of-ai-technologies-and-processingpersonal-information/.

Eldridge, J et al. (2023) "Cybersecurity and business interruption: foundations for prevention and mitigation", [online], https://www.lexology.com/library/detail.aspx?g=b2a3b8fb-e3dd-4987-9831-74a9934f2969&utm_source=Lexology+Daily+Newsfeed&utm_medium=HTML+email+-+Body+-+General+section&utm_campaign=ITechLaw+2015+subscriber+daily+feed&utm_content=Lexology+Daily+Newsfeed+2023-10-06&utm_term=.

ETsystems. (2023) "AI and privacy: The privacy concerns surrounding AI, its potential impact on personal data", [online], https://economictimes.indiatimes.com/news/how-to/ai-and-privacy-the-privacy-concerns-surrounding-ai-its-potential-impact-on-personal-data/articleshow/99738234.cms.

Goba, T. (2023) "Police violated POPI Act in Krugersdorp gang rape case – Information Regulator", [online], https://ewn.co.za/0001/01/01/police-violated-popi-act-in-krugersdorp-gang-rape-case-information-regulator

Illidge, M. (2023) "South Africa's leak and privacy watchdog cuts its teeth", [online], https://mybroadband.co.za/news/security/506966-south-africas-data-leak-and-privacy-watchdog-cuts-its-teeth.html.

Kahle, C. (2023) "Going beyond 'Nigerian Prince': SA turning into Africa's cybercrime capital", [online], https://www.citizen.co.za/lifestyle/technology/south-africa-turning-into-cybercrime-capital/.

Labuschagne, H. (2023) "South Africa's privacy watchdog investigating data breaches at credit bureaus", [online], https://mybroadband.co.za/news/security/507116-south-africas-privacy-watchdog-investigating-data-breaches-at-credit-bureaus-in-south-africa.html.

Legodi, S. (2023) "Cybercrime files: How AI is making virtual crimes easy and convincing", [online], https://ewn.co.za/2023/10/09/cybercrime-files-how-ai-is-making-virtual-crimes-easy-and-convincing.

Malinga, S. (2023) "POPIA principles must align with AI governance, say experts", [online], https://www.itweb.co.za/content/RgeVDvPRrn8MKJN3.

Mokdad, A et al. (2023) "Q&A: the data protection legal framework in South Africa", [online], https://www.lexology.com/library/detail.aspx?g=28b3e6c1-88d0-4cd4-8b33-4a9fbcfc0d15&utm_source=Lexology+Daily+Newsfeed&utm_medium=HTML+email+-+Body+-+General+section&utm_campaign=ITechLaw+2015+subscriber+daily+feed&utm_content=Lexology+Daily+Newsfeed+2023-09-20&utm_term=

Molosankwe, B. (2023) "Gauteng businessman gets 15 year sentence for stealing information from data firm Experian", [online], https://www.news24.com/news24/southafrica/news/gauteng-businessman-gets-15-year-sentence-for-stealing-infromation-from-data-firm-experian-20230330.

Moyo, A. (2023) "InfoReg slaps justice department with historic R5m fine", [online], https://www.itweb.co.za/content/o1Jr5MxPmm2MKdWL.

Mostert C. (2023) "Unprecedented cyberattacks target government entities", [online], https://www.politicsweb.co.za/opinion/unprecedented-cyber-attacks-target-south-african-g.

Mzekandaba, S. (2023) "Information watchdog sees data breaches notifications double", [online], https://www.itweb.co.za/content/j5alrMQAJOQMpYQk.

Ndlovu, (2023) " "Justice department takes Info Regulator to court over R5-million fine", [online], https://techcentral.co.za/justice-department-info-regulator-court/233071/.

O'Regan V. (2023) "India hits Daily Maverick with malicious cyberattack after report on Modi's tantrum", [online], https://www.dailymaverick.co.za/article/2023-08-23-india-hits-daily-maverick-with-malicious-cyberattack-after-report-on-modis-tantrum/.

Patterson, N. (2023) "What is cybersecurity and why is it important", [online], https://www.snhu.edu/about-us/newsroom/stem/what-is-cyber-security.

Reuters. (2023). "UK watchdog fines Equifax $13, 4 million for role in cyber breach" [online], https://www.reuters.com/technology/uk-watchdog-fines-equifax-134-million-role-cyber-breach-2023-10-13/.

Tredger, C. (2023) "Data breaches cost SA's financial sector R73m on average", [online], https://www.itweb.co.za/content/Olx4zMkazYQv56km.

Vermeulen, J. (2023) "Information Regulator tests its teeth – slaps Department of Justice with R5 million fine", [online], https://mybroadband.co.za/news/security/498859-information-regulator-tests-its-teeth-slaps-department-of-justice-with-r5-million-fine.html.

Watney, MM. (2022) "Cybersecurity threats to and Cyberattacks on critical infrastructure: a Legal Perspective", [online], https://papers.academic-conferences.org/index.php/eccws/issue/view/7.

PhD Research Papers

Managing Cyber Security Debt: Strategies for Identification, Prioritisation, and Mitigation

Christo Coetzer[1] and Louise Leenen[1,2]
[1]University of the Western Cape, Cape Town, South Africa
[2]Center for Artificial Intelligence Research (CAIR), Cape Town, South Africa

coetzercs@gmail.com
lleenen@uwc.ac.za

Abstract: This paper explores cyber security debt, a technical debt arising from unaddressed security vulnerabilities in an organisation's IT systems. These vulnerabilities accumulate due to resource limitations, time constraints, and expertise gaps, potentially leading to security breaches and data compromises. The paper outlines the cyber security debt management process involving identification, prioritisation, and mitigation strategies. Drawing parallels to financial debt, the authors emphasise the escalating risks of delaying cyber security debt repayment. The paper underscores the significance of diligent debt management in maintaining digital resilience and mitigating cyber threats. The increasing interconnectedness of systems and rapid software development has given rise to a hidden challenge known as cyber security debt. Cyber security debt is posed as a subset of technical debt, encompassing the accumulation of security vulnerabilities within an organisation's IT infrastructure and applications. Drawing a parallel between cyber security debt and its financial counterpart, the authors underscore the grave risks of deferring debt repayment. Just as financial debt accrues interest, unresolved security vulnerabilities compound over time, elevating the likelihood of breaches and data exposure. A poignant case study of the Equifax breach exemplifies the real-world consequences of neglecting security debt management. The failure to patch a well-known vulnerability led to a colossal data breach, highlighting the urgency of addressing security weaknesses promptly. Complex in nature, cyber security debt materialises when organisations fail to address vulnerabilities during various operational life cycles. These vulnerabilities might remain concealed within IT architecture, legacy code, or third-party libraries, posing a formidable challenge to detection and resolution. By understanding the parallels between financial and cyber security debt and proactively managing the latter, organisations can enhance their ability to safeguard against evolving cyber threats and maintain a robust security posture.

Keywords: Cyber security, Cyber security debt, Cyber security debt management

1. Introduction

This paper delves into cyber security debt management, presenting a novel framework designed to address this critical issue. This framework, referred to as the "proposed framework" throughout this paper, represents an innovative approach to tackling cyber security debt. It offers a fresh perspective, drawing on the principles of established methodologies such as Factor Analysis of Information Risk (FAIR) (*The FAIR Institute - Quantitative Information Risk Management*, n.d.), to comprehensively assess and manage cyber security debt.

In a rapidly evolving landscape where new threats constantly emerge, it is crucial to understand the concept of technical debt management and its relevance to the cyber security industry. Technical debt refers to the impact organisations create by taking shortcuts within their technical environment (Cunningham, 1993). Technical debt is mainly referred to within software development. It includes the build-up of poorly designed, unnecessarily complicated, or unwanted code or software, which are deficiencies in internal quality that make it harder to modify and extend a system further. This paper emphasises the need to identify, prioritise, and mitigate accumulated security vulnerabilities to minimise the risk of security breaches and data compromises.

We explore the intricacies of cyber security debt management, delving into the proposed framework and its application within the cyber security landscape. By adopting a methodology grounded in FAIR and considering the principles of technical debt management, organisations can enhance their ability to safeguard their digital assets and minimise the risks associated with unaddressed vulnerabilities. The evolution of cyber threats necessitates our strategies, and this paper strives to provide a forward-thinking approach to meet this challenge head-on.

The next section provides an overview of technical and cyber security debt management. In section 3 the authors present a unique cyber security debt management model and the paper is concluded in Section 4.

2. Technical and Cyber Security Debt Management

Initially, the technical debt metaphor was mainly used as a communication device - shifting the dialogue from a technical vocabulary to a financial language made discussions clearer and easier to understand for non-technical

people (McConnell, 2008). One of the most fruitful areas of achievement in technical debt research is technical debt identification, which is the first step in technical debt management (Seaman et al., 2012).

By incurring technical debt, organisations can trade off quality against productivity. An advantage of this includes short-term reduced maintenance time as well as cost. This, though, comes with the cost of additional work in the future, equivalent to paying interest on a debt. In this sense, technical debt as a term characterises the effect of delaying tasks, such as maintenance, on technical-focused projects. Nonetheless, technical debt is not the same as financial debt. The significant difference is that the interest associated with technical debt may or may not ever need to be paid off (Seaman et al., 2012). For example, it is not necessary to refactor a technical item that is overly complex if no further changes will be requested on that item in the future. This uncertainty differentiates technical debt from financial debt and further complicates the issue. Still, it also allows an organisation to leverage technical debt for their projects. Incurring technical debt has benefits but should be managed before a more significant, long-term cost is realised. Thus, organisations have to balance the costs and benefits of technical debt and make informed decisions on how technical debt should be prioritised and managed.

Cyber security debt is a subset of the above-explained technical debt. Cyber security debt refers to the accumulation of security issues or vulnerabilities in an organisation's IT infrastructure or applications due to various factors, such as limited resources, time constraints, or lack of expertise. These accumulated security issues can create significant technical debt, leading to security breaches and data compromises. Cyber security debt management encapsulates identifying, prioritising, and mitigating accumulated security debt. It involves assessing the organisation's current security posture, identifying security debt items, estimating the cost and effort required to address them, and prioritising the things based on their potential impact and likelihood of exploitation.

The concept of technical debt has been studied and applied in the software engineering domain, where it refers to the cost of future rework due to shortcuts taken during the development process. Researchers have recently proposed using the same concept to cyber security to manage security debt (Kruchten et al., 2012). Understanding and assessing technical debt is crucial. Dan Geer and Gunnar Peterson suggest using a Margin of Safety calculation to compare the 'book value' of an organisation's IT assets with the security controls and services used to defend them to determine its technical or security debt ratio (Geer & Peterson, 2014). Rather than facing forced repayment, an organisation should understand its obligation and implement processes to manage it, such as investing in managed services or taking out cyber insurance.

Short falling on financial debt payments results in more interest and extends the time to reach the principal. Fast falling on cyber security debt does not only delay future projects; it also results in the piling up of vulnerabilities, which places an organisation at greater risk of malicious cyber security exploits. The longer these security debts remain unpaid, the more interest they will accrue to address the shortcomings of the original investment.

An example of cyber security debt management can be found in the Equifax breach (2017). The organisation failed to patch a known vulnerability identified in Apache Struts, an open-source web application framework. Although the patch had been available for months, the organisation failed to deploy it, costing them a cyber security breach that compromised the personal data of more than 147 million people.

The accumulation of cyber security debt occurs when an organisation completes a project without addressing the identified weaknesses and vulnerabilities. Existing cyber security debt worsens if a project delays addressing an identified vulnerability due to pressured timelines. Accumulative cyber security debt can cause immense negative impact within an organisation. Organisations need to spend time and effort managing cyber security debt to minimise cyber security debt.

Cyber security debt is complex and is created by failing to address cyber security within the various organisational defined life cycles, such as project management or software development. As with its financial counterpart, it can be difficult to identify, often hidden deep in an organisation's IT architecture, legacy code, third-party libraries, and even the fundamental economic principles upon which some business models are based (Van Der Walt, 2018). In some cases, the complexity can be so great that the average business may be unable to determine where the interdependencies lie.

Several notable works have paved the way for a deeper understanding of these intertwined concepts in the technical and cyber security debt management landscape. Researchers have extensively explored the realm of technical debt, which initially emerged in software engineering, encompassing the costs incurred due to shortcuts during development. Technical debt principles have been transposed into cyber security, emphasising

identifying, prioritising, and mitigating security vulnerabilities to prevent potential breaches and data compromises (Lehman, 1979). Furthermore, influential studies have advocated adopting technical debt management practices within cyber security (Kruchten et al., 2012). This cross-disciplinary perspective has illuminated the parallel between technical and cyber security debt, underscoring the importance of addressing the accumulated vulnerabilities in an organisation's digital infrastructure. Additionally, foundational works in the field of Factor Analysis of Information Risk (FAIR) by Hubbard (Hubbard, 2020) have provided valuable insights into risk assessment and management, serving as a fundamental framework that forms a cornerstone for the Cyber Security Debt Management Model introduced in this paper. By building upon these established foundations and extending them into the unique domain of cyber security debt, we aim to contribute a practical approach for organisations to manage their digital risk landscape better.

3. The Cyber Security Debt Management Model

In this section, we introduce the Cyber Security Debt Management Model, a unique framework developed to address cyber security debt. This novel model is distinctly tailored to the evolving landscape of digital cyber security risk management. We emphasise the need for fresh, innovative approaches to confront the challenges posed by cyber security debt in today's interconnected digital world. The Cyber Security Debt Management Model represents a significant contribution, bridging the gap between technical debt management and cyber security. The model is a comprehensive and innovative framework that redefines how organisations approach and mitigate the ever-growing challenge of cyber security debt. By combining risk assessment, financial valuation, and strategic resource allocation, this model offers a systematic and quantifiable approach to managing cyber security risks and ensuring long-term digital resilience.

Figure 1 illustrates the process flow of the model and is discussed in the subsections that follow.

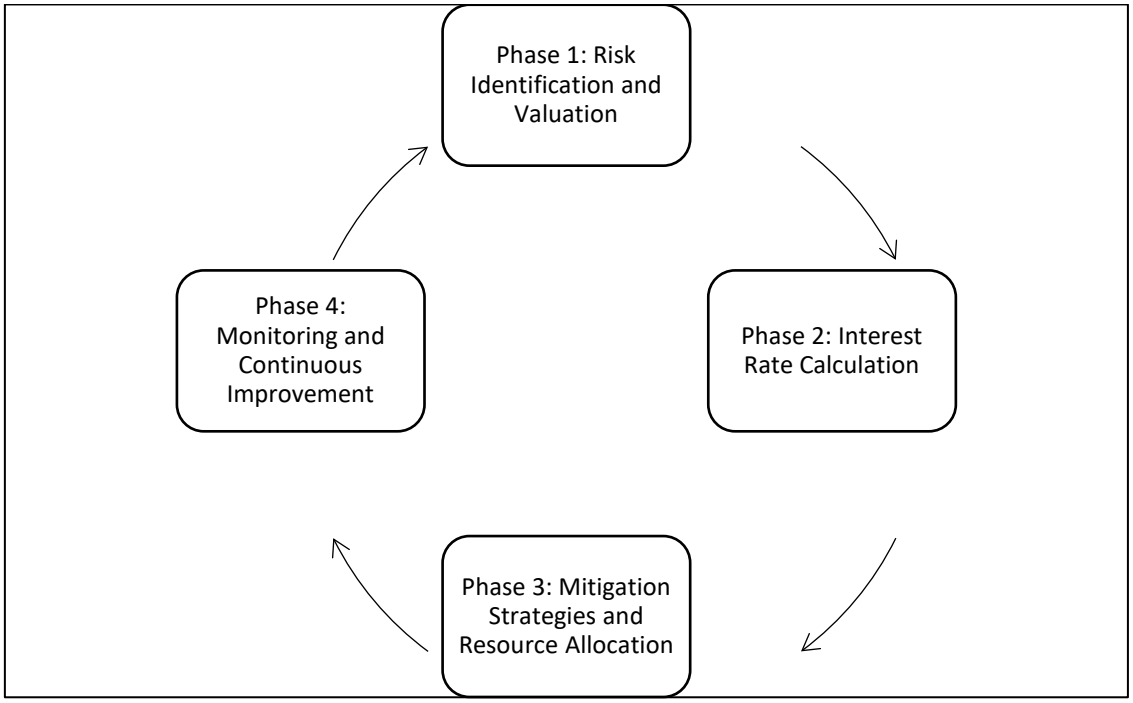


Figure 1: Cyber Security Debt Management Model (Own Compilation)

3.1 Risk Identification and Valuation

The cornerstone of the cyber security debt management model, the Risk Identification and Valuation component, offers a systematic approach to unearthing potential cyber security risks within an organisation's digital ecosystem. This phase is the foundation for effective risk management strategies in a landscape characterised by escalating threats and evolving vulnerabilities.

The process is aligned to the ISO/IEC 27005:2022 *Information security, cybersecurity and privacy protection — Guidance on managing information security risks* framework (ISO/IEC 27005, 2022) and commences with a meticulous and comprehensive identification of potential risks. This involves a rigorous analysis of an organisation's digital infrastructure, encompassing information systems, networks, software applications, and

critical assets. By scrutinising the entirety of the digital landscape, the model ensures that no potential risk goes unnoticed. These risks extend beyond technical vulnerabilities to threats posed by external actors, internal vulnerabilities, and emerging attack vectors.

Once identified, each risk undergoes a quantitative valuation to assess its potential financial impact. This valuation relies on a combination of factors, including the possible loss magnitude and the likelihood of risk occurrence. Each risk is assigned a numerical value that reflects its financial implications, allowing for an objective comparison and prioritisation of risks. By quantifying risks in monetary terms, organisations gain a pragmatic understanding of the potential consequences of cyber security debt. The quantification process not only aids in informed decision-making but also provides a crucial metric to communicate risk severity to non-technical stakeholders. Translating technical jargon into financial terms bridges the communication gap between cyber security professionals and organisational leadership. Decision-makers can comprehend the financial implications of unmitigated risks, facilitating more meaningful discussions and collaborative efforts toward risk mitigation.

The Risk Identification and Valuation component is a roadmap for organisations to address cyber security debt proactively. It enables organisations to identify potential vulnerabilities, quantify their impact, and lay the groundwork for effective resource allocation. This phase is instrumental in prioritising mitigation efforts, ensuring that resources are allocated to areas with the highest potential for risk reduction. As the digital landscape evolves and new risks emerge, the component's iterative nature allows organisations to regularly reassess and recalibrate their risk valuations, ensuring continued alignment with the ever-changing threat landscape.

In essence, the Risk Identification and Valuation component of the cyber security debt management model empowers organisations to unearth potential vulnerabilities systematically, assign them tangible financial values and lay the foundation for a comprehensive risk mitigation strategy. This component equips organisations with the insights needed to navigate the intricate cyber security landscape with resilience and efficacy by quantifying risks and fostering a holistic understanding of their financial implications.

FAIR is a robust and widely adopted framework for quantifying and analysing information and cyber security risks (*The FAIR Institute - Quantitative Information Risk Management*, n.d.). Developed by the Open Group, FAIR provides a structured methodology that enables organisations to assess and prioritise risks based on quantitative measurements, enhancing decision-making and resource allocation in cyber security. The framework incorporates financial concepts and risk analysis principles to understand risk exposure comprehensively. The following actions are prescribed:

3.1.1 Integrating FAIR into "Risk Identification and Valuation"

Integrating FAIR principles enhances the "Risk Identification and Valuation" phase in the cyber security debt management model context. By incorporating FAIR into this phase, organisations can enrich their risk assessment process with a quantitative perspective, aligning seamlessly with the model's objective of quantifying risk exposure and prioritising mitigation efforts. FAIR introduces several vital components that can be integrated into the Risk Identification and Valuation phase:

Risk Taxonomy and Definitions: FAIR emphasises the importance of a standardised risk taxonomy and precise definitions for various risk factors. Incorporating FAIR's risk categories and descriptions augments the accuracy of risk identification, ensuring a consistent and shared understanding among stakeholders.

Loss Event Frequency (LEF) and Loss Magnitude: LEF and loss magnitude are fundamental concepts in FAIR. Incorporating these concepts into the model's risk identification phase allows for a more structured assessment of how frequently specific loss events may occur and the potential financial impact if they materialise.

Scenario Analysis: FAIR encourages the creation of well-defined scenarios to assess risks comprehensively. This aligns with the model's objective of scenario design and execution. By integrating FAIR's scenario analysis approach, organisations can delve deeper into risk identification and valuation, capturing nuanced risk profiles.

Risk Measurement: FAIR employs calibrated estimation techniques to quantify risk factors. This can be applied to estimate probabilities and potential loss magnitudes associated with various risks. Integrating FAIR's risk measurement techniques enhances the model's ability to accurately quantify interest rates and cyber security debt.

Data Collection and Analysis: FAIR places emphasis on data collection and analysis to inform risk assessments. Integrating FAIR's data-driven approach into the Risk Identification and Valuation phase ensures that evaluations are grounded in empirical evidence, enhancing the accuracy and reliability of risk valuations.

The cyber security debt management model leverages a quantitative and standardised framework for assessing and quantifying risks by infusing FAIR principles into the Risk Identification and Valuation phase. This integration facilitates a more rigorous and comprehensive understanding of risk exposure, enabling organisations to prioritise mitigation efforts based on quantifiable data and aligning with the overarching goal of effective cyber security debt management.

FAIR enriches the Risk Identification and Valuation phase by introducing structured risk taxonomy, quantitative measurement techniques, scenario analysis, and data-driven assessments. By incorporating FAIR principles, organisations can enhance their ability to identify, assess, and quantify risks accurately, ultimately strengthening the foundation of the cyber security debt management model's risk assessment process.

3.2 Interest Rate Calculation

A defining innovation within the cyber security debt management model, the Interest Rate Calculation component, introduces a dynamic and intuitive mechanism to quantify the potential financial consequences of unmitigated cyber security risks. Unfortunately, there were not many new research papers found in the field of cyber security debt with a focus on interest calculations. Drawing parallels from the financial domain, where interest rates reflect the cost of borrowing money, this component leverages this concept to express the cost of deferring risk mitigation efforts over time. In this context, the interest rate encapsulates the cumulative financial impact that a particular cyber security risk may incur if left unaddressed. Computed through the fusion of risk occurrence probability and potential loss magnitude, the interest rate paints a vivid picture of the urgency and severity associated with each risk. A high-interest rate signifies a rapidly accumulating cyber security debt, spotlighting the criticality of timely risk mitigation.

This quantitative representation of risk underscores the far-reaching implications of cyber security debt. Like compounding interest in finance, unmanaged risks can compound over time, magnifying financial repercussions. The interest rate is a decision-support metric and offers a tangible measure to illustrate the potential long-term costs of neglecting risk mitigation efforts.

The interest rate calculation introduces a uniform framework for risk prioritisation, allowing organisations to compare and rank risks based on their calculated rates objectively. This empowers decision-makers to allocate resources prudently, directing them towards addressing risks with the highest potential financial impact. In doing so, the organisation can optimise resource utilisation, strategically deploying mitigation measures to areas that promise the most significant reduction in cyber security debt. Furthermore, the interest rate calculation component transcends technical jargon, enabling effective communication between cyber security professionals and non-technical stakeholders. By expressing cyber security risks in terms familiar to financial discussions, the model facilitates a shared understanding of risk severity. This alignment enhances communication, enabling decision-makers to grasp the financial stakes and make informed choices regarding resource allocation and risk management strategies. In addition, the Interest Rate Calculation component provides a dynamic and adaptable metric that evolves with the ever-changing cyber security landscape. As threat vectors evolve and organisational priorities shift, the model accommodates adjustments to risk occurrence probabilities and potential loss magnitudes. This flexibility ensures that the calculated interest rates remain relevant and reflect emerging risks, empowering organisations to respond proactively.

The Interest Rate Calculation component of the cyber security debt management model presents a ground-breaking methodology to quantify the financial implications of cyber security risks. Drawing inspiration from financial principles and applying them to cyber security, this component offers a tangible and intuitive metric that enhances risk prioritisation and fosters effective communication. It empowers organisations to address cyber security debt with strategic insight. The model highlights the intersection of financial and cyber security considerations through the interest rate calculation, forging a path towards proactive risk mitigation and sustainable digital resilience.

3.3 Mitigation Strategies and Resource Allocation

The Mitigation Strategies and Resource Allocation component is a pivotal pillar within the cyber security debt management model, translating quantitative risk assessments into actionable strategies that fortify an

organisation's cyber security posture. This dynamic phase harnesses the insights from risk valuation and interest rate calculation to guide judicious resource allocation and strategic risk mitigation.

With interest rates that quantify the potential financial consequences of unaddressed cyber security risks, supported by debt calculation, organisations embark on a strategic journey of resource allocation. The calculated interest rates serve as a compass, directing decision-makers towards risks with the highest potential impact. By focusing resources on threats with elevated interest rates, organisations ensure that their mitigation endeavours yield maximum value in risk reduction and cyber security debt alleviation.

The component's essence lies in optimising resource utilisation to address the most pressing vulnerabilities. It empowers organisations to allocate personnel, time, and financial investments where they are most needed. High-interest-rate risks demand immediate attention and a robust allocation of resources to mitigate their potential financial ramifications. This strategic deployment of resources enables organisations to enact proactive measures that reduce the long-term accrual of cyber security debt.

Mitigation strategies span a spectrum of approaches, from technical solutions such as software patching and network segmentation to non-technical measures like employee training and incident response planning. The model facilitates a comprehensive understanding of the most effective strategies for each risk, resulting in a tailored and pragmatic approach to risk mitigation. Resources are channelled strategically, reflecting an organisation's unique risk profile and priorities.

The component also instils a culture of accountability by establishing measurable targets for risk reduction. Progress is tracked through a tangible decrease in interest rates, reflecting the diminishing potential financial impact of mitigated risks. This quantifiable improvement enables organisations to showcase their commitment to cyber security resilience, bolstering stakeholder confidence and demonstrating proactive risk management.

Furthermore, the iterative nature of the model ensures that mitigation strategies remain adaptable to the evolving threat landscape. As new risks emerge and organisational priorities shift, the component enables organisations to recalibrate resource allocation and adjust mitigation strategies accordingly. This dynamic responsiveness is crucial for maintaining an effective and up-to-date cyber security posture, safeguarding against emerging vulnerabilities.

The Mitigation Strategies and Resource Allocation component empower organisations to leverage quantified risk assessments for strategic decision-making. Organisations take decisive steps toward minimising cyber security debt by allocating resources where they are most impactful and tailoring mitigation strategies to individual risks. This component fosters a culture of proactive risk reduction, enhancing digital resilience and positioning cyber security as an integral aspect of organisational sustainability. Organisations forge a path toward a secure digital future through strategic resource allocation, adeptly managing the cumulative consequences of cyber security debt.

3.4 Monitoring and Continuous Improvement

The bedrock of the cyber security debt management model's resilience and effectiveness lies within the Monitoring and Continuous Improvement component. In an era marked by dynamic and relentless cyber threats, this phase serves as a strategic compass that enables organisations to navigate the ever-evolving digital landscape with agility and foresight.

This component recognises that cyber security risks are not static entities but dynamic forces that demand ongoing vigilance. It underscores the importance of continuous monitoring, regular reassessment, and iterative refinement. As the threat landscape mutates and organisational priorities shift, the model adapts to ensure that risk assessments remain relevant and resource allocation remains aligned with emerging risks. By consistently monitoring the efficacy of risk mitigation strategies, organisations can gauge the actual impact of their efforts. The model encourages a proactive approach, allowing organisations to identify potential gaps, emerging vulnerabilities, or changes in risk profiles. This real-time assessment empowers organisations to refine their strategies promptly, ensuring that cyber security debt remains under control and that the risk landscape is met with adaptive defences.

The iterative nature of the component extends to the interest rate calculation, which allows organisations to adjust risk valuations and potential loss magnitudes based on evolving circumstances. This flexibility ensures that the model remains responsive and adaptable to new threat vectors, technologies, and organisational developments. As a result, the calculated interest rates accurately reflect the potential financial impact of cyber security risks, guiding resource allocation with precision. Moreover, the continuous improvement cycle

reinforces a culture of proactive risk management and accountability. Stakeholders witness tangible progress as interest rates decrease over time, showcasing the organisation's commitment to reducing cyber security debt. This transparent and quantifiable improvement strengthens stakeholder trust and fosters a culture of collective responsibility, positioning cyber security as a fundamental aspect of overall business resilience.

The Monitoring and Continuous Improvement component also holds the potential to uncover valuable insights that inform strategic decision-making. Organisations can glean actionable intelligence by analysing trends and patterns in risk valuations and mitigation efforts. These insights empower organisations to forecast potential risk trajectories, allocate resources pre-emptively, and identify emerging threats before they escalate into substantial cyber security debt.

The Monitoring and Continuous Improvement component of the cyber security debt management model ensures the model's relevance and efficacy in the face of an ever-changing threat landscape. By instilling a proactive mindset, fostering adaptive risk management, and providing a mechanism for ongoing assessment and refinement, this component enables organisations to mitigate cyber security risks with agility and foresight. It underscores the importance of sustained vigilance, resilience, and continuous evolution in pursuing long-term digital security and mitigating cyber security debt.

4. Conclusion

In the face of the relentless evolution of digital landscapes and the pervasive threats they bring, adopting the cyber security debt management model emerges as a seminal advancement in cyber security risk management. This innovative model encapsulates a transformative approach that transcends traditional boundaries, offering organisations a robust framework to combat the burgeoning challenge of cyber security debt effectively.

The model's emphasis on quantification presents a paradigm shift that addresses the limitations of qualitative risk assessments. By translating cyber security risks into tangible financial terms, decision-makers gain unprecedented clarity into the potential consequences of unmitigated vulnerabilities. This newfound comprehension enhances communication between technical experts and non-technical executives, fostering critical collaborative efforts in an increasingly interconnected business environment.

Integrating financial concepts within cyber security signifies a monumental leap towards aligning security efforts with overarching business objectives. The model's ability to articulate risks in financial language bridges the gap between cyber security and executive leadership, positioning cyber security debt as a strategic concern integral to an organisation's overall risk management strategy. This alignment ensures that cyber security professionals and executives have a common understanding, enabling well-informed decision-making that resonates with broader business goals.

The dynamic nature of the model, underscored by the iterative process of continuous monitoring and improvement, amplifies its resilience in the face of ever-evolving threats. As the threat landscape mutates and organisational priorities shift, the model adapts by recalibrating risk valuations and resource allocations. This agility ensures that organisations remain proactive in addressing emerging vulnerabilities, thereby pre-emptively curbing the accrual of cyber security debt.

The cyber security debt management model engenders a culture of accountability, optimisation, and resilience within organisations. The model enables organisations to proactively reduce cyber security debt and curtail potential financial losses by quantifying risks, determining interest rates, and steering resource allocation. It positions cyber security as an imperative facet of strategic decision-making, emphasising the importance of strategic resource allocation and fostering a proactive stance against the long-term implications of unmanaged risks. As organisations confront an increasingly intricate digital landscape, adopting the cyber security debt management model offers a formidable arsenal for risk mitigation. Its integration of financial and cyber security concepts and its quantifiable framework empowers organisations to address cyber security debt comprehensively. In doing so, the model equips organisations to navigate the complexities of the digital age, safeguard valuable assets, and ensure sustained success in an evolving and interconnected world.

References

Cunningham, W. (1993). The WyCash portfolio management system. *ACM SIGPLAN OOPS Messenger*, *4*(2), 29–30. https://doi.org/10.1145/157710.157715

Equifax. (2017). *Equifax Releases Details on Cybersecurity Incident, Announces Personnel Changes | Equifax*. https://investor.equifax.com/news-and-events/press-releases/2017/09-15-2017-224018832

Geer, D., & Peterson, G. (2014). Measuring Security Book Value. In *Columns - Margin of Safety or Speculation* (pp. 49–51).

Hubbard, D. W. (2020). The failure of risk management: Why it's broken and how to fix it. *The Failure of Risk Management: Why It's Broken and How to Fix It*, 1–366. https://doi.org/10.1002/9781119521914

ISO/IEC 27005. (2022). *ISO/IEC 27005:2022 Information security, cybersecurity and privacy protection — Guidance on managing information security risks*. https://www.iso.org/standard/80585.html

Kruchten, P., Nord, R. L., & Ozkaya, I. (2012). Technical debt: From metaphor to theory and practice. *IEEE Software*, *29*(6), 18–21. https://doi.org/10.1109/MS.2012.167

Lehman, M. M. (1979). On understanding laws, evolution, and conservation in the large-program life cycle. *Journal of Systems and Software*, *1*(C), 213–221. https://doi.org/10.1016/0164-1212(79)90022-0

McConnell, S. (2008). *Managing Technical Debt | Construx*. https://www.construx.com/resources/whitepaper-managing-technical-debt/

Seaman, C., Guo, Y., Zazworka, N., Shull, F., Izurieta, C., Cai, Y., & Vetrò, A. (2012). Using technical debt data in decision making: Potential decision approaches. *2012 3rd International Workshop on Managing Technical Debt, MTD 2012 - Proceedings*, 45–48. https://doi.org/10.1109/MTD.2012.6225999

The FAIR Institute - Quantitative Information Risk Management. (n.d.). Retrieved August 16, 2023, from https://www.fairinstitute.org/

Van Der Walt, C. (2018). *Understanding and Assessing Technical Debt for Improved Cybersecurity*. https://www.cbronline.com/opinion/technical-debt-cybersecurity

An AI Model for Digital Forensic Readiness in the Cloud Using Secure Access Service Edge

Taurai Hungwe[1] and Hein Venter[2]
[1]Department of Computer Science and Information Technology, Sefako Makgatho Health Sciences University, South Africa
[2]Department of Computer Science, University of Pretoria, South Africa

taurai.hungwe@smu.ac.za
hventer@cs.up.ac.za

Abstract: Computing infrastructure has evolved and has brought about changes to the ways that work and business are carried out. Cloud computing has redefined these workspaces by providing connectivity and tools to enable productivity, collaboration, and flexibility. As work moves outside the centralised office and goes remote, users are accessing the cloud directly, leaving the protection of the corporate network and leaving the users' computing platforms open to threats. In the pre-cloud era, the data centre for organisations was the single location where digital assets would be housed and non-complicated security parameters implemented. The firewall would be the main security perimeter implemented to secure the network in this pre-cloud era. The advent of cloud computing has brought potential areas and gaps in securing the organisational data, information and communication connectivity to the cloud-based resources. As such, there is need to rethink and redesign the models which can be implemented to secure the cloud computing services. The cloud should be in a state of digital forensics readiness in order to facilitate digital forensics investigation. The study focuses on the development of an artificial intelligence model for digital forensic readiness for the cloud using secure access service edge. This integrated approach might assist in the provisioning of cloud security.

Keywords: Digital forensics readiness, Cloud computing, Artificial intelligence, Secure access service edge, Security

1. Introduction

The current working environments have been redefined as organisations embrace the use of cloud computing. The use of cloud computing services results in an increase in remote users and computers. Cloud computing provides connectivity and tools to enable productivity, collaboration, and flexibility (Cisco, 2022). The recent Covid-19 pandemic has also accelerated the adoption of the cloud (Islam et al., 2021). Employees, users and computers connect remotely to the cloud and leave the secure corporate network environments as work is conducted outside the centralized office. This increased the acceptance of the cloud, however, it exposes the organisations and users to security threats and attacks. A need has arisen for a rethink and redesign of security models to be implemented to secure data, information, networks and computers accessing the cloud.

In the day-to-day provisioning of cloud services, security breaches might occur. For example, unauthorised data access by cybercriminals to sensitive personal information and data may occur such as unauthorised access to people's financial records stored on the cloud. When such security infringements occur, an investigation is required to investigate the security transgressions. Digital forensics investigation (DFI) is a process of recovering and investigating digital evidence. The gathered evidence might be used to present arguments in a court of law or used to reconstruct events on a computing system.

In order to facilitate the DFI process, specifically in the cloud, certain controls and processes should be in place prior to the occurrence of a security breach. The cloud should have been made ready for digital evidence collection in some way. Thus, in order to do this, the cloud components must apply digital forensic readiness. Tan (2001) explained digital forensic readiness (DFR) as positioning organisations to have proactive digital forensic (DF) techniques in place. This optimises the collections of evidence to meet legally reliable standards and reduces the overall effective costs associated with DFI processes. Cloud services bring in aspects of volatility. For example, some cloud service providers do not provide persistent storage to virtual machine instances (Purnaye & Jyotinagar, 2015). At any given time or moment, a cloud service might be available and the same service might not be available to the users or organisation in – often soon – subsequent time. Security breaches could have occurred, at the time the cloud service was available to the user and that same service might not be available when a DFI processes would need to be conducted. This volatility on cloud services provisioning poses challenges in carrying out DFI processes as some or all of the data might be lost by the time of investigation. Again, for a successful DF process to be conducted, the cloud components and services must be digital forensically ready. Currently, cloud computing services which are made available or accessed by users through computing edge (referring to a range of networks and devices at or near the user) are not digital forensically ready and not much research has been conducted in this area (MacDonald et al., 2019; van der Walt & Venter,

2022). This paper focuses on security challenges of the cloud and comes up with a model using AI for digital forensics readiness. The objectives are, (1) Identify the security challenges of cloud computing edge components and evaluate if the components and services are digital forensically ready (DFRy); and (2) Design an artificial intelligence (AI) framework for digital forensic readiness for the core components for secure access service edge (SASE). For this, AI unsupervised algorithms should be applied to detect security threats and attacks on the cloud offerings. Secondary data is used for the unsupervised algorithm for learning, testing and then finally for predictions on cloud computing edge core components for DFR.

Piecemeal security offerings might provide security to different cloud services. Such siloed approach might be cumbersome and not being the best way to provide cloud security. Available frameworks on securing cloud components are relatively new and little research has been conducted (MacDonald et al., 2019; van der Walt & Venter, 2022). In year 2022 there were only three (3) relevant articles collectively dealing with the concepts of security, DF and DFR in relation to cloud computing (Van der Walt & Venter, 2022). One of the research gaps identified by Van der Walt and Venter points to the non DFR on cloud computing core components. A model is proposed to achieve DFR in the cloud. Moreso, without an AI model on DFR for cloud computing, it remains costly and time consuming to find out what would have been the security transgressions when a DFI process is conducted. It will assist in making cloud computing components DFRy and align the model to existing models and standards.

The rest of this paper is structured as follows: Section 2 presents the background of the study. Section 3 is a contribution to the study which proposes the first version of the model. Section 4 critically evaluates the proposed model, and then section 5 concludes the paper and outlines future research work.

2. Background

This section discusses relevant concepts related to the research. These concepts are, defining cloud computing, benefits of cloud computing, cloud computing challenges, cloud services and deployment models, edge components and cloud security issues.

Cloud computing (CC) is the on-demand access of computing services over the Internet (Baror et al., 2021; IBM, 2023; NIST, 2015). The computing services include, infrastructure, platform and software provisioned through cloud service providers (CSPs). For example, software applications (MS 365) as a service over the internet with the software hosted by a CSP as opposed to owning and licensing of the software on the organisational (user) premises. Such services can rapidly be provided, scaled up or scaled down depending on demand and usage with very little effort and user interaction (Armbrust, et al., 2010; Armbrust et al., 2009; Maun, 2018). One of the main advantages of CC is that the user pays for the usage (pay-as-go) of the software service, but not for the actual software itself with the user particularly owning the software. However, CCS are dependent on the availability of the internet, without internet connectivity services cannot be accessed.

Accessing of the cloud is through the use of edge components also referred to as endpoint or end things. For example, end points include, routers, gaming controllers, vehicles, cameras, laptops, personal computers (PCs), internet of things (IoT) and mobile devices Thus, the edge component are virtual or physical devices fixed at the end (edge) of the network (Yang, et al., 2020). With many devices connecting to the cloud, concerns on security in the cloud arise.

Cloud security involves the securing of cloud computing systems (Kaspersky, 2022). Cloud security is made up of three categories, namely, data security, identity and access management (IAM) and governance. Data security involves the protection of data (such as data stored in a database) against destruction, unauthorised access and data breaches among others. IAM involves granting of correct rights to authorised users to computing resources. Governance provides the requisite policies, processes and procedures to control, direct and manage security risks, threats, and attacks for an organisation. Security in cloud computing caters for the network, data, data storage, virtualisation frameworks, operating systems (OS), middleware, runtime environments, applications and end-user hardware.

This includes keeping data private and safe across online-based infrastructure, applications, and platforms Cloud security is shared between the cloud user and the CSP. The CSP is responsible for the security of the cloud. Security of the cloud entails safety and awareness of the cloud users in the process of using CCS. The security of the cloud includes hardware, software, networking, and facilities that run the cloud services. The user is responsible for the security in the cloud. This implies that, the user manages security aspects on data, classifying assets, access, and cloud configurations. For example, consider the management of and use of user passwords. Security in the cloud deals with the securing of the cloud in the running of applications, storing data and processing transactions. In other

words, it deals with the safety of the cloud itself. Security in the cloud encompasses ways and technology (tools and machinery) that protects cloud computing environments against cybersecurity threats.

The main advantages or benefits to businesses of CC are the increase of efficiency, productivity, flexibility and scalability. CCS is scalable and flexible since it enables any organisation to set up and manage information technology (IT) infrastructure on a pay-as-you-use basis, enabling cost effective data redundancy, off-site backups, big data storages and IT sourcing (Sarga, 2012). However, despite the many advantages of CC, the cloud also provides an environment for cybercriminals to conduct illegal activities, for example, distributing malware, conducting scams, identity theft and other criminal activities (Pasquale, et al., 2016; Pichan et al., 2015). Criminal activities leave digital footprints behind which can be collected and analysed through DFI process. DFI is presented in the next section.

Digital forensics is part of forensic science which assists in the recovery, preservation, analysis and presentation of digital evidence. Digital evidence can be used in legal evidence processes and assists to posit motive behind an incident with intention to identify the offender.

Several definitions on digital forensics exist in literature (Agarwal et al, 2011; Cohen et al., 2011; Ray, 2009; Vacca, 2005; Gary, 2001). Digital forensics investigation (DFI) is a process of securing, collection, preservation, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purposes of deriving scientific evidence of criminal activities is conducted (Gary, 2001). The following definition captures the core of concept and processes involved in digital forensics. Digital forensics is "the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal or helping to anticipate unauthorized actions shown to be disruptive to planned operations" (Gary, 2001). An undisputed digital forensic process must be conducted which stands any scrutiny against the law and evidence presentations. Harmonised Digital Forensic Investigation Process (HDFIP) follows applicable standards to conduct DFI process (ISO/IEC 27043, 2015) and will be used in this study. Without having the CC edge components DFR, it is difficult to conduct a DFI.

Digital forensics readiness (DFR) is "the state of the organisation where controls are in place, in order to facilitate the digital forensic processes and to assist in the anticipation of unauthorised actions shown to be disruptive to planned operations" (Pangalos & Katos, 2010). DFR is a proactive way of making sure that there is an ability to carry out a credible DFI when an incident occurs. To achieve DFR, Rowlingson (2004) suggested ten activities to be done before an incident occurs. These ten activities are as follows:

1. Define the business scenarios that require the collection of digital evidence.
2. Identify available sources and different types of potential evidence.
3. Determine the evidence collection requirement(s).
4. Establish a capability for securely gathering legally-admissible evidence to meet the requirement(s).
5. Establish a policy for secure storage and handling of potential evidence.
6. Ensure monitoring is targeted to detect and deter major incidents.
7. Specify circumstances when escalation to a full, formal investigation should be launched.
8. Train staff in incident awareness, so that all those involved understand their role in the digital evidence process and the legal sensitivities of evidence.
9. Document an evidence-based case describing the incident and its impact.
10. Obtain legal review on the incident. This ensures that the evidence collection process is forensically and legally correct and the collected evidence is sufficient in order to link the offenders with the committed crime for prosecution purposes.

The implementation of such activities prepares organisations for carrying out DFIs with minimum costs and less disruption of business continuity in a proactive manner for the cloud. Given the nature of cloud computing being volatile (at any given time a cloud computing service would be available to the cloud user and shortly thereafter might not be available), time is of essence for conducting a successful DFI. Thus, this can be achieved by having the cloud being DFRy. Considering Rowlingson's ten activities to achieve DFR mentioned earlier on, such activities could be automated through artificial intelligence (AI). Definition, use and role of AI in DFR follows.

AI involves process automations of learning, self-correction and reasoning in order to do human activities through the use of computing machines and technologies. For example, AI can be used to conduct and anlyse financial credit ratings of an individual based on the financial behaviour of the individual. AI is, therefore, the process of mimicking human intelligence in machines (Dunsin et al., 2022; Agarwal & Karahanna, 2000). AI replicates or simulate human intellect, such as learning and reasoning with accelerated speed, with intensified

power, and with increased complexity. This study will not dwell much on the history of AI except for a few notable developments in AI. This is presented in the next paragraph.

AI dates back to science, philosophy and history. The development of AI is owed to Alan Turing in 1950. Coining of the term was done by John McCarthy in 1956 at a conference in Dartmouth College (Collins et al., 2021). John McCarthy defined AI as "the science and engineering of making intelligent machines" (McCorduck, 2004). It was during this time that the birth of AI was realised, termed and defined. In defining AI, it is worthy to split the terms to artificial and intelligent, define the two terms, form an understanding of the artificial and intelligence terms and render a combined definition for AI. Trocin et el. (2021) considers intelligence as a way of sense making using of information amassed from previous encounters and dealing with improbabilities of future or new activities. The term artificial is the imitation of human-like cognitive tasks with clear ways of doing things (Benbya et al., 2021). Merging the definitions of the two terms, the following definition emerges which is adopted in this research. AI is defined as a way in which machines use computational power to collect, discover, infer and learn from data to accomplish programmed activities to impersonate humans (Mikalef & Gupta, 2021). Having defined AI, the next aspects is to consider the use and role of AI.

The role of AI is to provide computing machines that can decipher data and perform human-like activities. The human-like activities are performed due to various reasons, for example, a machine can be used to perform an activity that – being performed by a human – could expose the human to hazardous environmental factors. In some instances, AI is used to automate repetitive activities which are easily programmable on a machine, but physically too exhausting for a human. Process, product and service improvements are enabled through the use of AI. Furthermore, AI enables improvement and refinement of algorithms, thereby increasing the accuracy of results when these algorithms are implemented. AI also have self-correcting mechanisms to improve processes and activities. In this study, the role of AI is directed towards digital forensics readiness (DFR) within the cloud, as described in the next section.

DFR, as earlier defined, is the ability to respond proactively and collect digital evidence related to a security incident with minimal cost or interruption to services (Yasinsac & Manzano, 2001; Dunsin et al., 2022). DFR processes might be improved through automation using AI, as the DFR processes are currently conducted manually. For example, AI can be used to predict the occurrence of security incidents in order to provide anticipated responses to reduce security attacks. AI has the potential to engage in multi-step reasoning, to understand the meaning of natural language, to design innovative artefacts, to generate novel plans that achieve certain goals, and even reason about its own reasoning (Langley, 2019). It is through these possibilities that AI can be exploited to untangle security incident reports in order to prepare computing infrastructure to become DFRy.

To assist in the automation of digital forensics investigations, Rughani (2021) came up with an AI-based digital forensics framework which automates major routine operations in digital forensic investigations using intelligence acquired from artificially-trained data. This is an example of process automation and optimisation to specifically increase the efficiency in a specific application area of AI. The future of AI can also be linked with this research focusing on an AI framework for digital forensic readiness for secure access service edge components. This might transform the manner in which digital forensics readiness (DFR) would be applied on the edge components in cloud computing.

3. AI Model for Digital Forensic Readiness in the Cloud using Secure Access Service Edge

Users access the cloud by means of edge components; see Figure 1. Examples of edge components include, among others, mobile devices such as mobile phones, gaming controllers, water sensors, vehicles, cameras and PCs. An AI model is implemented for DFR to enhance cloud security using secure access service edge (SASE).

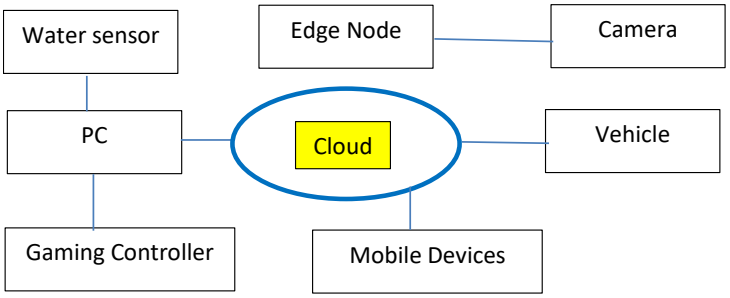


Figure 1: Relationship of the cloud and edge components

Figure 2 illustrates the proposed AI model. The deployment of SASE is implemented in the edge components and the cloud. This is the new contribution to making cloud computing edge components to be digital forensically ready. The model proactively implements SASE and monitors the components and gathers potential digital forensic evidence (PDFE). At the edge side, the security breaches are detected, and remedial actions are conducted. Similarly, on the cloud side, the SASE is deployed.

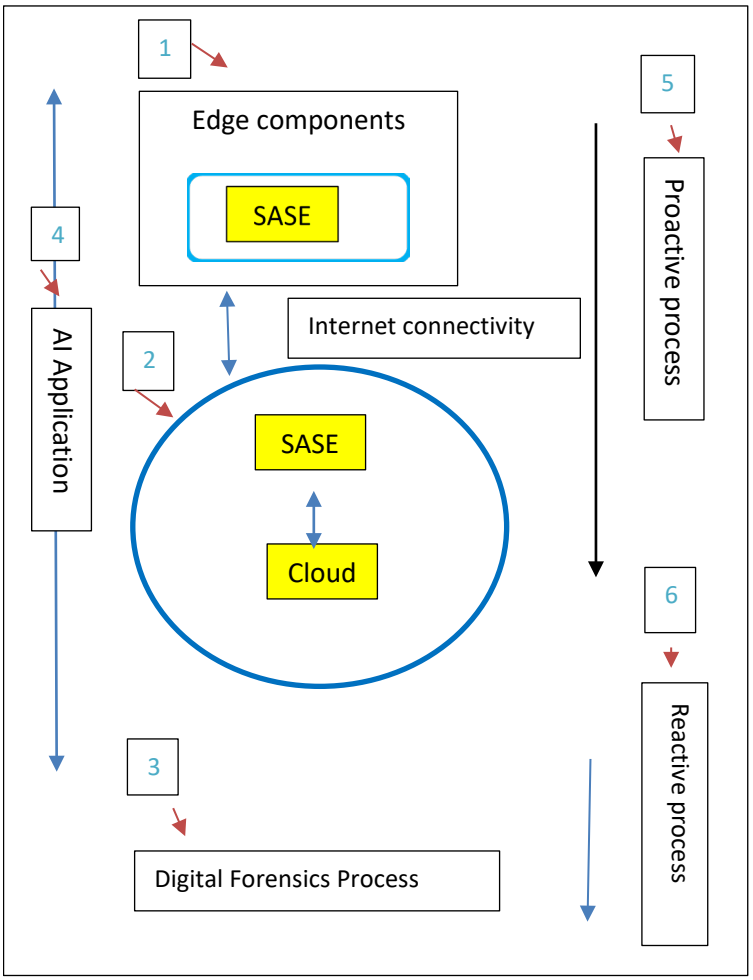


Figure 2: Proposed AI DFRy models

Incoming network traffic from the edge to cloud components and outgoing traffic from the cloud to the edge components is monitored for security breaches. The aspects of detecting the security breaches in the edge components and the cloud could be automated using AI. The AI automated processes feed into the parallel proactive processes of DFRy. The proactive processes approach is based on preparing for DFR based on Rowlingson's (2004) ten activities which should be done before an incident occurs, active monitoring, gathering and retaining digital information within the cloud environment and the edge components. The DFI process is a reactive process, meaning it is conducted after a security incident has occurred. It is then after a security breach has happened that a DFI process is initiated as contrasted to proactive DFR process. The main role of the model is to have the edge components and the cloud to be DFRy. This assists when a need arises to conduct DFI processes. The detailed proposed AI DFRy model is illustrated in the next paragraph.

Annotation "1", in Figure 2, shows how the edge components such as mobile devices access the cloud services. SASE is implemented at the edge components. Mobile devices have the potential to be attacked by malware and introduce the malware to the network, to other connected devices and to the cloud. Thus, the edge components are the entry point of malicious activities which compromise the other components and the rest of the cloud environment. The other entry points would be the network and the cloud. Monitoring of these entry points is implemented and PDFE is also gathered at these edges of the cloud. PDFE is stored in the cloud infrastructure such as storage and servers. Storage and servers will preserve the PDFE to be used for DFR purposes. On the cloud, at annotation "2", SASE is also implemented. Similarly, PDFE is gathered, stored and preserved in the

cloud infrastructure. DFI processes, at annotation "3", are conducted using the information proactively collected, stored and preserved in instances of security breaches.

The AI application, denoted by annotation "4" illustrates the implementation of unsupervised machine learning unsupervised learning algorithms. The algorithms will detect the security threats and incidences. For example, if there is an upsurge of unexplained traffic on the network connecting the edge components and the cloud, it might signify a security incidence. The AI application will also automate the activities proposed by Rowlingson as ways to prepare the cloud and the edge components to be DFRy. The proactive process, at annotation "5", indicates the manner in which the edge components and the cloud are prepared for the next process, i.e. the reactive process at annotation "6" of carrying out a DFI. Annotation "6" is instituted after a security incidence would have been detected, hence it is referred to as a reactive process. The proactive process is implemented in anticipation of a security incidence. Thus, making sure that cloud environments and the edge components are able to collect PDFE, stored in appropriate storage spaces and preserved. To illustrate the DFR process, a run through Rowlingson's activities follows.

Assume that in a business setting, for example, a mobile device such as a mobile phone connects to the cloud through a public cloud service provider. This provides the first activity of defining the business scenario that necessitates the collection of digital evidence. The identification of available sources and different types of PDFE then follows, which is the second activity under Rowlingson's activities. In the process of identifying the phone and possibly the owner, certain PDFE can be identified such as the phone's international mobile equipment identity (IMEI), mobile station international subscriber directory number (MSISDN), subscriber identity module (SIM), international mobile subscriber identity (IMSI) number. More PDFE can be identified but is omitted here due to space constraints. Activity three, determination of the evidence collection requirement(s), is then conducted. This would involve activities such as commissioning of the investigation by the responsible person and identifying the scope of the investigation. Thereafter, the data and information must be stored and preserved. Thus, activity four, establishes a capability for securely gathering legally-admissible evidence to meet the requirement(s) of the mobile phone data. The mobile phone data collected in this scenario will have to be securely stored and handled according to established policies. Activity six then ensures that monitoring is targeted to detect and deter major incidents in the use of the mobile phone. If unwarranted mobile phone usage is detected, then the next activity, activity seven, has to commence. In activity seven, circumstances are specified when such unwarranted mobile phone usage is escalated to a full, formal investigation would be launched. The training of staff in incident awareness, so that all those involved understand their role in the digital evidence process and the legal sensitivity of evidence, commences in activity eight. Staff are sensitized on matters such as privacy and observing legislative instruments, for example, in South Africa such matters are protected by the protection of personal information act (POPI Act) of 2013 (Government gazette, 2013). In activity nine, documentation of the evidence-based case describing the incident and its impact is put in place. The last activity (activity ten) is obtaining legal guidance of the incident. This ensures that the evidence collection process is forensically and legally correct and the collected evidence is sufficient in order to link the offenders with the committed crime for prosecution purposes. All these activities are there to have the cloud and the connected edge components of the cloud prepared for an eventuality that PDFE can be collected in the event that a DFI has to be conducted. The next section presents the critical evaluation of the model.

4. Critically Evaluation of the Proposed Model

The proposed AI model is a novel concept that automates the activities in the cloud and its edge components by gathering digital information and preserving it forensically in preparation for DFI purposes. This is the contribution of the research which aims at forensic planning and preparation for a DFI process within this particular environment. The AI model identifies and predicts the security challenges of cloud computing edge components and applies unsupervised algorithms in the detection of security threats and attacks. The information detected on incidences prepares the cloud and the edge components to actively collect PDFE. By making the edge components and the cloud ascribe to Rowlingson's activities, the cloud and the components are able collect credible evidence before an incident occurs, thus, making them digital forensically ready (DFRy). The model enables cloud and the edge components to achieve sufficient forensic preparedness from the forensically preserved information by using SASE and implementing Rowlinson's ten readiness activities. Implementing such activities, if an incident is detected, assists the users using the cloud to gather credible forensic evidence that can stand legal tests. The model simplifies the process of evidence collection as required data and information are easily collected from the digital forensically ready environment. The next section concludes this paper.

5. Conclusion

The increase in the use of cloud computing (CC) has also increased the number of connected edge devices and remote users. With such an increase, the wider CC environment is littered with more security risks. Security incidents in the course of using the cloud require DFI processes to unearth the security breaches. Efficient DFI are premised by how digital forensically ready (DFRy) the devices are. Cloud computing edge components are not DFRy. Without having the DFR components in place, it is difficult to conduct a DFI. The research focuses on making the cloud computing components and services DFRy and proposes an AI model for DFR in the cloud using SASE. The proposed model was illustrated as it deploys integrated SASE and edge components in the cloud. AI is then applied to identify, predict and report security incidences. The reports are then used to prepare the cloud edge components for DFR. The proactive preparations for the edge components provide for an efficient and cost-reduced DFI process. This achieves the aim of the research for putting forensic preplanning processes in place for the proactive preparation of a DFI.

References

Agarwal, A., Gupta, M., Gupta, S., & Gupta, S.C. (2011). Systematic Digital Forensic Investigation Model. *International Journal of Computer Science and Security (IJCSS),* Vol. 5 No. 1, 2011, pp. 118 – 131.

Agarwal, R., & Karahanna, E. (2000). Time Flies When You're Having Fun: Cognitive Absorption and Beliefs about Information Technology Usage. *MIS Quarterly*, 24(4), p.665.

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee., G, Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. A view of cloud computing. (2010). Commun ACM. 2010;53:50–58. doi:10.1145/ 1721654.1721672.

Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M. (2009). Above the clouds: A berkeley view of cloud computing. Department of Electrical Engineering and Computer Science. Univ. California, Berkeley, Rep. UCB/EECS; Pg1-26, 2009. Online available at http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.html Last Accessed 10 February 2024.

Baror, S.O., Venter, H.S. & Adeyemi, R. (2021). A natural human language framework for digital forensic readiness in the public cloud. Available at: https://repository.up.ac.za/bitstream/handle/2263/78877/Baror_Natural_2021.pdf?sequence=1&isAllowed=y> Last accessed 12 April 2023.

Benbya, H., Pachidi, S., Jarvenpaa, S.L. (2021). Special issue editorial: artificial intelligence in organizations: implications for information systems research. *J. Assoc. Inf. Syst.* 22, 281–303.

Cisco. (2020). How cloud security reduces threat risk and paves the way to SASE. Modern solutions for protecting the new way businesses work. Available at https://umbrella.cisco.com/info/ebook-how-a-cloud-security-solution-reduces-risk-and-paves-the-way-to-sase. Accessed 19 February 2023.

Collins, C., Dennehy, D.; Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda*. International Journal of Information Management* 60 (2021) 102383. https://doi.org/10.1016/j.ijinfomgt.2021.102383.

Cohen, F., Lowrie, J., & Preston, C. (2011). *The State of the Science of Digital Evidence Examination.* In: Peterson, G. and Shenoi, S. (Eds), Advances in Digital Forensics VII. Cohen, F. (2012). Digital Forensic Evidence Examination. (4ed), California: Fred Cohen & Associates

Dunsin, D.; Ghanem, M. & Ouazzane, K. (2022). The use of artificial intelligence in digital forensics and incident response (DFIR) in a constrained environment. August 2022. *International Journal of Information and Communication Technology* 16(6):280-285. Available at: https://repository.londonmet.ac.uk/7708/3/Australia_Dip_paper.pdf. Accessed 24 June 2023.

Gary, P. (2001). A Road Map for Digital Forensic Research, Report from DFRWS 2001, *First Digital Forensic Research Workshop*, Utica, New York, August 7 – 8, pp. 27–30.

Government Gazette, REPUBLIC OF SOUTH AFRICA. Vol. 581 Cape Town 26 November 2013. /https://www.gov.za/sites/default/files/gcis_document/201409/3706726-11act4of2013protectionofpersonalinforcorrect.pdf. Accessed 11 December 2023.

IBM. (2023). IBM X-Force threat intelligence index 2023 report. Available at: www. https://www.ibm.com/reports/threat-intelligence. Accessed 12 June 2023.

Islam, M.N., Colomo-Palacios, R., & Chockalingam, S. "Secure Access Service Edge: A Multivocal Literature Review," 2021 21st International Conference on Computational Science and Its Applications (ICCSA), Cagliari, Italy, 2021, pp. 188-194, doi: 10.1109/ICCSA54496.2021.00034.

ISO/IEC 27043. (2015). *Information technology – Security techniques – Incidents investigation principles and process.* Geneva: ISO

Kaspersky. (2022). What is Cloud Security? Available at: https://www.kaspersky.com/resource-center/definitions/what-is-cloud-security (Accessed: 27 May 2022).

Langley, P. (2019,). An integrative framework for artificial intelligence education. *In Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 9670-9677).

MacDonald, N., Orans, L., & Skorupa, J. (2019). *'The Future of Network Security Is in the Cloud'.* Gartner, (August).

Maun, U. (2018). Future of Cloud Computing 2025: Trends & Predictions. [Online]. https://www.seasiainfotech.com/blog/history-and-evolution-cloud-computing/. Last Accessed 10 February 2024.

McCorduck, P. (2004*). Machines who think: A personal inquiry into the history and prospects of artificial intelligence (2nd ed.).* A. K. Peters.

Mikalef, P., & Gupta, M. (2021). Artificial Intelligence capability: conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance*. Inf. Manage.* 58, 103434 https://doi.org/10.1016/j.im.2021.103434.

NIST. (2015). *The NIST Definition of Cloud Computing.* Available at: https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-145.pdf. Accessed 12 April 2023.

Pangalos, G., & Katos, V. (2010). Information Assurance and Forensic Readiness. *e-Democracy*, 181-188.

Pasquale, L., Hanvey, S., McGloin, M., Nuseibeh, B., Pasquale, L., & Hanvey, S. Adaptive evidence collection in the cloud using attack scenarios. *Computer Security.* 2016;59:236–254. doi:10.1016/j.cose.2016.03.001.

Pichan, A., Lazarescu, M., Soh, S.T., Cloud forensics: technical challenges, solutions and comparative analysis. *Digit Investigation*. 2015;13:38–57. doi:10.1016/j.diin.2015.03.002.

Purnaye, P., & Jyotinagar, V. (2015). Cloud forensics: Volatile data preservation. *International Journal of Computer Science Engineering (IJCSE).* ISSN: 2319-7323 Vol. 4 No.02 March 2015, pp 41 – 43.

Ray, C. (2009). *Distributed Database Systems*. New Delhi: Pearson Education.

Rowlingson, R. (2004). "A ten step process for forensic readiness". *International Journal of Digital Evidence,* 2(3), 1-28.

Rughani, P.H. (2021). Artificial Intelligence Based Digital Forensics Framework*. International Journal of Advanced Research in Computer Science.* Volume 8, No. 8, September-October 2017. DOI: http://dx.doi.org/10.26483/ijarcs.v8i8.4571.

Sarga, L. (2012). Cloud Computing: An Overview. *Journal of systems integration* (2010). 3 (4), 3–14.

Tan, J. (2001). *Forensic readiness.* Cambridge, MA: @ Stake, 1-23.

Trocin, C.; Våge Hovland, I., Mikalef, P., & Dremel, C. How Artificial Intelligence affords digital innovation: A cross-case analysis of Scandinavian companies. *Technological Forecasting & Social Change* 173 (2021) 121081.

Vacca, J. (2005). *Computer Forensics: Computer Crime Scene Investigation. (2ed*). Massachusetts: Thompson Course Technology

Van der Walt, S.P., & Venter, H. (2022). Research Gaps and Opportunities for Secure Access Service Edge. *Proceedings of the 17th International Conference on Information Warfare and Security,* 2022, pp 609 – 619.

Yang, C., Lan, S., Wang, L., Shen, W., & Huang, G. G. (2020). Big data driven edge-cloud collaboration architecture for cloud manufacturing: a software defined perspective. *IEEE access, 8,* 45938-45950.

Yasinsac, A., & Manzano, Y. (2001). Policies to Enhance Computer and Network Forensics*. Proceedings of the 2001 IEEE Workshop on Information Assurance and Security*. United States Military Academy, West Point, NY, June 2001.

On the Benefits of Vulnerability Data Consolidation in Application Security

Santanam Kasturi[1], Xiaolong Li[2], Peng Li[3] and John Pickard[3]
[1]Technology Management, Indiana State University, Terre Haute, USA
[2]Dept. of Electronics and Computer Engineering, Indiana State University, Terre Haute, USA
[3]Dept. of Technology Systems, East Carolina University, Greenville, USA

skasturi@sycamores.indstate.edu
xiaolong.li@indstate.edu
lipeng@ecu.edu
pickardj@ecu.edu

Abstract: This research aims to build upon a conceptual idea of consolidating all application security vulnerability data from monitoring, detection, and discovery tools into a physical system that allows for convergence of observation and response to an event that is a threat. Multiple application security testing and monitoring tools are deployed at different layers of an application architecture and capture activities that occur at that layer. This multi-layer data capture is disconnected without any analysis of data lineage from the externally exposed web attack surface to deep down into the application and data layers. It is only through this data consolidation can one provide a reliable statistical analysis of correlating multiple vulnerability information and synthesize an attack pattern and predict possible events accurately. The benefits of such a system are discussed in this paper that includes how one can organize the data, identifying temporal and spatial correlation of events, focusing on specific web requests that point to a specific vulnerability, and formulating a fast response to such events. Advantages of integrating with Security Information and Event Management (SIEM), Security Orchestration, Automation, and Response (SOAR/XSOAR), Extended Detection Response (XDR) are briefly discussed. The analysis can be further used to develop a predictive system using deep learning (DL) techniques using correlation of application security vulnerability information.

Keywords: Data Consolidation, Attack Surface, SIEM, XSOAR, SOAR, XDR, Vulnerability Correlation, Deep Learning

1. Introduction

According to a survey, 77% of CISO's report that prioritization of vulnerability management remains a challenge as risk assessment is slow and lacks enough information, while another 88% say that combining application runtime context, vulnerability analysis, and risk impact assessments would make the job easier (Dynatrace, 2023). A typical 'breakout time' - the time an attacker takes to move laterally from the first penetration point in a compromised host to another host within the system - declined from 98 minutes in 2021 to 84 minutes as measured in 2023 (CrowdStrike, 2023). If one does not respond with an action to stop the spread to minimize the costs of damages caused by hackers, within this breakout time of 84 minutes, the loss could be catastrophic especially in the case of a ransomware attack. It is recommended that a 1 - 10 - 60 rule be established where the first minute should be in detecting the threat, the next 10 minutes be spent in understanding the threat, and the final 60 minutes in responding to the threat with appropriate action to minimize losses (CrowdStrike, 2023). Even more critical is a zero-day vulnerability commonly found in open-source software libraries and components which is a software code available to all to use and share updates and helps developers to use code developed by others for certain features that are readily available and can be used with little or no modifications. These components are again small functions or features that can be used among multiple applications within an organization. Chances are that everyone using a version of a component will introduce the same vulnerability across the enterprise, as was seen with the Log4Shell, a zero-day vulnerability that was discovered in 2021, and one that had the potential for an exploit that had no grace period to remediate, and the exposure was instantly open to attackers, leaving 'zero-days' to fix the vulnerability.

2. Challenges with Application Security

Multiple challenges exist and three are discussed here. By multiple challenges we mean there are operational challenges in integrating a host of tool technologies, managing information gathered from each, maintaining data integrity and quality, producing metrics and key performance indicators, and presenting visualizations for leadership are many and are not discussed here. However, it is worth mentioning that there are challenges beyond the four described below.

2.1 Application Security Testing in a Siloed Manner

Due to the nature of the applications that are being developed and approaches to different test and scan types that include Static Analysis Security Testing (SAST), Software Composition Analysis (SCA), Dynamic Analysis Security testing (DAST), Application Ethical Hack (AEH) /Pen test, Application Programming Interface (API) testing, Container testing, and even Infrastructure as Code scanning (IaC), in most large organizations data consolidation is not integrated to study correlation of vulnerabilities end-to-end and analyze if there is a pattern to attacks happening from outside. One solution proposed discusses elimination of false positives by running correlations from vulnerabilities found from different test and scan types within an application, all housed within one tool (Checkmarx, 2022). This however does not resolve the challenge of external attacks that target a specific vulnerability within an application or even a collection of applications that serve a single business domain that have common nodes of transactions. We must also consider that some organizations employ external agencies to scan their exposed web attack surface to discover vulnerabilities that are not detected by any of the test types. This can be due to a new vulnerability that has been added by National Vulnerability Database (NVD), or due to missed coverage in case of risk-based testing. So, other capabilities that discover vulnerabilities that complement those found using the standard test and scan methods employed within, must be included to cover the entire attack surface. Unless one brings in all kinds of application security vulnerabilities detected by internal tests, external agency discoveries, and external monitoring by Web Application Firewall (WAF) to consolidate all vulnerability data it will be difficult to visualize or synthesize an attack pattern (Checkmarx, 2022; Akamai, 2023; Carielli et al, 2022).

2.2 Security Testing Tool Sprawl

As more and more organizations adopt tools and automation, the abundance of tools in the market can lead to multiple security testing solutions that have over the years become outdated with the ever-growing complexity of applications. Added to this challenge is the amount of data that comes out of all this testing and the analytics still relying heavily on spreadsheet-based skills and reporting. The tool sprawl, semi-automated security testing and reporting practices makes analysis of vulnerability information difficult, slow, and is a challenge to proactively respond to security incidents. With the advent of AI and ML one can hope for an autonomous analytics solution, however the challenge here is data quality. Studies show that 40% of challenges come from tool sprawl, and correlating alerts from different tools is largely manual and time consuming, resulting in too many false positives. These studies also show that 62% of organizations use four or more solutions to maintain the security of their applications (Dynatrace, 2023). Deploying the right monitoring and detecting tool becomes key to correlate attack requests with existing vulnerability information as shown in Table 1. The details of each type of analysis and the tools associated with it can be found in literature referred here (Carielli, et al, 2021; Checkmarx, 2022; Veracode, 2023; Primeon, 2018; Akamai, 2023; Signal Sciences, 2021; FASTLY, 2022; Na, 2021; SALT, 2023)

Table 1: Test / Scan Data

Capability	Deployment Layer	Vulnerability Information	Alerts
SAST	Code	CWEs, Severity, Exploitability	Point in time scan reports
SCA	Component	CVEs, Severity, Exploitability	Point in time scan reports
API SECURITY	API	CVE's, CWE's, Severity, Exploitability	Point in time reports
DAST	URL/UI	CWEs, Severity, Exploitability	Point in time scan reports
PENETRATION TEST	URL/UI	CWEs, Severity, Exploitability	Point in time test reports
WAF	Network Layer	Blocked Attack requests, Incoming threats, target IP and host information	Continuous monitoring
Hybrid WAF/ RASP	Application Layer	Blocked Attack requests, Incoming threats, target IP and host information	Continuous monitoring

2.3 Data Storage

A final challenge is storing such a vast amount of information from multiple sources of data with relational capabilities enabling rapid analysis and visualization. Data consolidation from multiple sources that reveal vulnerabilities at different layers of an application becomes a critical requirement to begin with. Without it, to think about identifying an attack pattern, synthesizing an attack, or even predicting and preventing an attack will be fraught with risks of too many false positives.

3. Data Consolidation Benefits

In view of the three challenges discussed above - application testing happening in a siloed manner, tool sprawl, and data storage - the need for an integrated platform that consolidates data from monitoring and detecting tools and efficiently manages and provides analytical insights is much desired.

The first step towards a predictive analytics system is to build a platform for data consolidation from all the different monitoring and detecting methods employed for application security testing. The final solution is to design an integrated analytics and predictive system using autonomous processing techniques, like a Deep Learning (DL) technique such as an artificial neural network (ANN) that will predict and preempt an attack and is beyond the scope of this work and requires a separate treatment. Creating a data warehouse that can support data storage from multiple sources and ensure providing data marts for individual lines of business use is essential. The information from all the test, scan, and monitoring results provides insight of vulnerability and attack information as various attributes as seen in Table 1. To understand and correlate the data, one must look at a fast and easy way to store and retrieve information that can easily be fed to a visualization tool like Power BI or Tableau. The same information also needs to be fed to a predictive system that runs an algorithm for DL and assumes reliability and accuracy. The entire flow of data and the system is shown in Figure 1.

Combining all types of vulnerability information in a single data warehouse and transforming the data into visualization models to be presented using a BI Analytics solution has specific advantages. The first advantage is the ability to correlate vulnerabilities discovered in multiple layers through multiple monitoring and detection capabilities. Presenting vulnerability information discovered in multiple layers provides a view to the association of vulnerabilities to the transaction path.

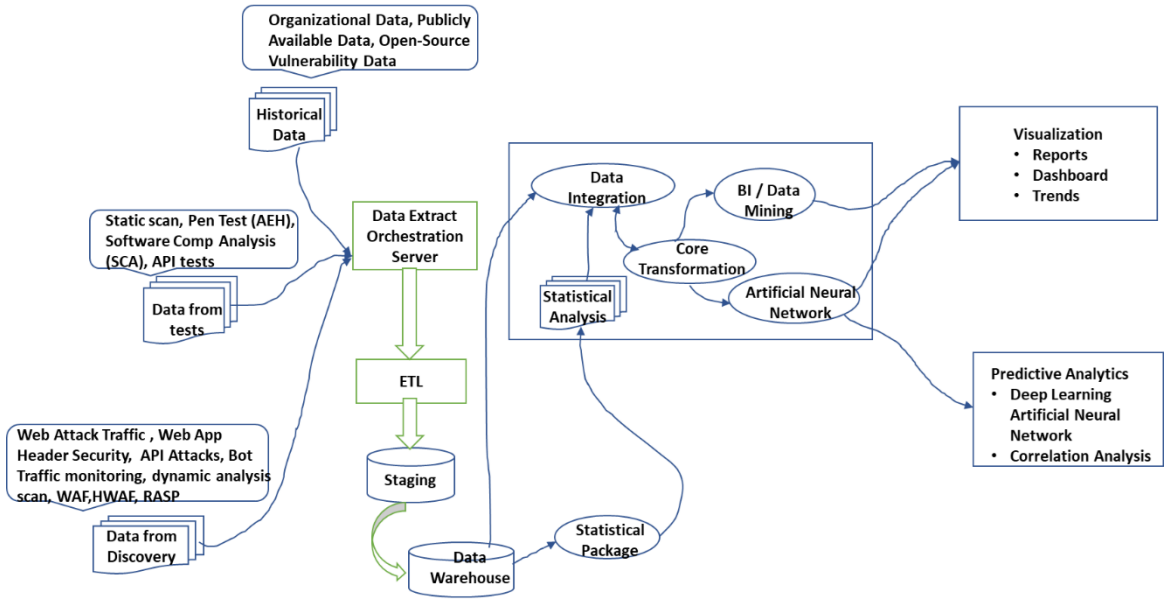


Figure 1: Conceptual Integrated Data Analytics System

It still does not give you a full picture of the threats facing your attack surface. It still does not tell you the present and the future; all it tells you is the past. While data consolidation will help in understanding the vulnerability lineage, it still will not help you in responding with the speed needed in the event of an attack. It will require a continuous monitoring system, and combined with all the data gathered will provide a better picture of what may happen. This continuous monitoring tool can be a WAF, or a Hybrid WAF, or even better a RASP tool (Akamai, 2023), (Carielli et al, 2022), (Signal Sciences, 2021), (FASTLY, 2022), (Na, 2021). In our research we have

used a WAF. Further, this can be integrated with an incident response system like SIEM/XSOAR/XDR that are primarily designed for networks, endpoints, and sensors, and not ideally for applications.

3.1 Correlation Analysis

Here is why data consolidation is key because processing all such correlations, which may be hundreds, is humanly impossible and to provide immediate response is not easily realizable, and not without a lot of manual analysis. Here is also why AI and DL are necessary to effect autonomous processing and replace manual processing and make decision support systems meaningful. Analyzing requests monitored using WAF and vulnerability data from SAST, SCA, DAST scans combined with a penetration test, one finds correlations between types of attack requests and corresponding vulnerabilities found in the application (Kasturi, et al, 2023). It is not possible, when dealing with dozens of applications with transactions running into several thousands in each, to manually identify what is a normal transaction and what is not. A funnel down approach is needed which becomes part of an extended analysis. Once specific request types that are considered as valid and applications with corresponding vulnerabilities are identified, analysis can be done at an application-level correlating with specific vulnerabilities to study if they are indeed valid requests or missed attacks. This requires a deeper analysis of the data and is part of continuing research to provide key metrics and a possible lead up to synthesizing an attack pattern by studying inter-application correlations. This analysis becomes a key piece in the integration with incident management and response systems as discussed in the following sections.

3.2 Integrating SIEM/XSOAR/XDR

When it comes to security and compliance, organizations must keep track of a vast array of data sources and threats. This is where SOAR (Security Orchestration, Automation, and Response), XSOAR (Extended Cybersecurity Orchestration, Automation, and Response), SIEM (Security Information and Event Management), and XDR (Extended Detection and Response) come into play. SIEM is a technology that is used to collect, store, and analyze security data from various sources. The data is then used to detect and respond to security threats. SIEM technology can be used with other security solutions, such as firewalls, antivirus software, and intrusion detection systems for detecting and alerting anomalies. SOAR on the other hand is a system for automating, orchestrating, and responding to security threats after collecting data from same sources that the SIEM gathers. These responses can be customized to specific threats. If the data is not primed for an immediate response, it can be used for further analysis to make a more accurate response. XSOAR is an extended improvement over a SOAR in that it incorporates additional features like machine learning (ML), threat intelligence (TI), and security analytics for improved automation and orchestration capabilities, with a dashboard for unified security information for better threat response. XDR combines multiple detection and response techniques like SIEM and XSOAR into a single platform, combining data from different sources (network traffic, endpoint logs, and threat intelligence) to provide a more comprehensive picture of security events. Security teams gain speed through this technology integration in responding to threats by getting a full view of the security posture. XDR is not a replacement to SIEM or XSOAR but acts as a complement to them.

A survey of application of SIEM/XSOAR/XDR technologies in a variety of industries is discussed briefly. The literature covered starts with challenges identified (Cinque, et al, 2018) to the latest review on the state of SIEM (Velasquez et al, 2023). The industries and domains covered are Air Traffic Control (ATC) (Cinque, et al, 2018), Intrusion Detection Systems (IDS) (Muhammed, et al, 2023), Industrial Control Systems (ICS) and critical infrastructures (Mern, et al, 2022; Gonzalez-Granadillo, Gonzalez-Zarsosa and Diaz, 2021), and Wind Energy Systems (Johnson, et al, 2023). This gives a good sweep of systems that are not just IT, but also includes critical infrastructures outside of IT. In addition to SIEM, the literature also includes scope of XSOAR, EDR, and XDR. Tables 2 - 5 below gives a comprehensive summary of SIEM, SOAR/XSOAR, EDR, and XDR (Nour, Pourzandi and Debbabi, 2023; Olteanu, 2022).

Table 2: Security Information and Event Management (SIEM)

Characteristics	Advantages	Challenges	Deployment
Collect security event logs and telemetry data in real time for threat detection and compliance use cases, analyze data in real time, analyze incidents and impact, report, store logs and relevant information.	Irreplaceable, provides a holistic view, establishes a threshold for critical control, provides a data lineage to trace back to the initial attack and ensures	Vulnerable to attacker countermeasures, expensive to deploy and maintain, correlation to attack source and target is a challenge, can generate a lot of alerts and false	Network, devices, sensors, and all infrastructure components leading to event collection; event normalization; set action rules for protect, remove, and respond; event storage and monitoring.

	hardening of the security posture and provides a sure means for real-time analysis. Supports audit and mandatory regulatory requirements in managing threats.	positives, requires skilled analysts, attack variants can pose problems, a new and evolving market with too many players, scalability is a big factor and has a multiplier effect and with an ever-growing number of connected devices and assets leading to alert fatigue with little context to the security operations team to understand and implement a response.	

Table 3: Endpoint Detection and Response (EDR)

Characteristics	Advantages	Challenges	Deployment
Endpoint monitoring and collecting potential threat activity, analyzing data, and identifying malicious patterns, provide automatic response with action to stop or remove threats with alerts.	EDR's can provide critical context to detect advanced threats, can run automated response activity to isolating an endpoint from the network in real-time, to stop and prevent further spreading of the issue.	Standardization of unstructured data; setting correlation rules are business specific and depends on analyst knowledge; behavior analysis is based on knowledge of system, environment, and a-priori knowledge; requires extensive ground truth data to create behavior rules.	Resources located on the endpoint like collected events, logs, and binaries, are correlated, and analyzed to determine whether a suspicious activity occurs on the host, are signature-based solutions for pattern recognition. Does not cover networks.

Table 4: Extended / Security Orchestration, Automation and Response (X/SOAR)

Characteristics	Advantages	Challenges	Deployment
X / SOAR systems (Extended / Security Orchestration, Automation, and response can be used to collect data on information security events from multiple monitoring targets, process them, and configure an automated response using typical response scenarios, can be run as a playbook.	Plays together with SIEM, uses threat and telemetry data across a wide range of target monitoring systems to identify events and alert threats. Automating routine and repeatable incident response tasks and workflows	Playbooks can become repetitive and if not reviewed can become obsolete. Although they save a lot of time in response actions, they are created by humans from experience and a-priori knowledge of systems. As systems and attacks change, playbooks need to be revisited.	Host based intrusion detection systems (HIDS) and Network based intrusion detection system (NIDS) are used to create playbooks for responses based on what is detected. Streamlines incident response through an interface.

Table 5: Extended Detection and Response (XDR)

Characteristics	Advantages	Challenges	Deployment
Extended Detection and Response is a technology for threat detection and response, it unifies security products for detection and response and threat intelligence into a single platform.	Uses machine learning, correlation, and analytics capabilities to enhance the response time and the efficiency of the security teams. While SIEM creates alerts, XDR does a deeper analysis using AI/ML for a better representation of the threat.	It is a new concept; the technology is evolving and has not yet matured.	Covers endpoints, servers, emails, cloud, and networks.

3.3 Monitoring: Application Security vs Infrastructure Monitoring

In recent times, web applications have become the primary targets of attackers with their widespread use that even attackers identified as script kiddies with little knowledge can perform very sophisticated and damaging

web attacks using ready-to-use attack tools. In addition, according to the analysis performed on all vulnerabilities, the rate of critical vulnerabilities in the web application layer is approximately four times that of the network layer (Edgescan, 2019). Yet, most literature today reports studies on network-based intrusion detection systems (NIDS) than studies on web-based attack (Sevri and Karacan, 2019).

Applications have “normal” attributes and behavior. This is the “descriptive” portion of the evolutionary state of analysis (Descriptive, Diagnostic, Predictive, Prescriptive, and Cognitive). Attributes are scripts, languages, interpreters, compilations, configurations, security artifacts, connections (to other apps or databases), and infrastructure topology. Behavior points to availability, stability, response time, throughput/volumes, utilization (hard disk, memory, CPU), load conditioned activities (batch vs. pseudo-batch), workflow, error rates, and error response activities (or lack thereof) and indicates a state of normal behavior or a state of a compromised system. Basically, the artifact that defines a workflow profile would be considered an attribute, while the actual processing statistics of a transaction through the workflow would describe a set of behaviors (Kasturi, 2020). Monitoring can be at the infrastructure or at the application level. While network and device monitoring are in the domain of infrastructure monitoring, network traffic associated with a specific application is in the domain of application monitoring. Similarly, CPU utilization or heap size measure is in the domain of infrastructure monitoring but associating that to a business transaction from an application is in the domain of application monitoring. The challenge, however, is that monitoring tools do not provide the correlation between an infrastructure analysis system behavior to an analysis of application behavior. The reason is applications, and their transactions do not reside or are contained in one system in an enterprise that has distributed application architecture. Applications are multi-tiered and are spread across vast global geographies. An application is also not a network, device, or a host, it is a piece of code that is not a physical entity for an agent to monitor and report to an SIEM, and when we have hundreds of distributed applications and many talking to each other, and all having hundreds of weaknesses it can be visualized as a spaghetti code with trapped vulnerabilities. There is a disconnect between a security operations infrastructure and application behavior, and correlating both needs to be addressed to bring in application monitoring within the ambit of security operations and threat response using SOAR or XSOAR, or even XDR.

3.4 Integrating Application Security Vulnerability Information Data Platform with SIEM/XSOAR/XDR

Incident response using SIEM/XSOAR/XDR are all reactive systems, these come into play after the fact that a threat has been detected. This is where threat hunting combined with threat detection complemented with historical data on existing application vulnerabilities and technical debt plays a pivotal role in making a comprehensive system. The proposed data consolidation platform is the first step towards that as it enables creating a data warehouse producing datamarts, dashboards, and a decision support system. The data, however, becomes a static representation of vulnerabilities prevailing in the application code, and this must be used with a correlation analysis (Kasturi et al, 2023). The end goal is to integrate this system into an incident response system as shown in Figure 2. The correlation analysis data can be fed into the incident response platform using STIX and TAXII to be analyzed as a threat pattern. Structured Threat Information Expression (STIX) is a standardized language that uses a JSON-based lexicon to express and share threat intelligence information in a readable and consistent format across indent response systems. Trusted Automated Exchange of Intelligence Information (TAXII) is the format through which threat intelligence data is transmitted. TAXII is a transport protocol that supports transferring STIX insights over Hyper Text Transfer Protocol Secure (HTTPS). Conversion of vulnerability data into a reliable threat indicator requires analysis and understanding, and interpretation of the vulnerability and the associated risk prior to creating a threat indicator. Once the feeds are ingested into the incident response platform, while SIEM creates alerts, XDR does a deeper analysis using AI/ML for a better representation of the threat, and X/SOAR runs a playbook for automated response.

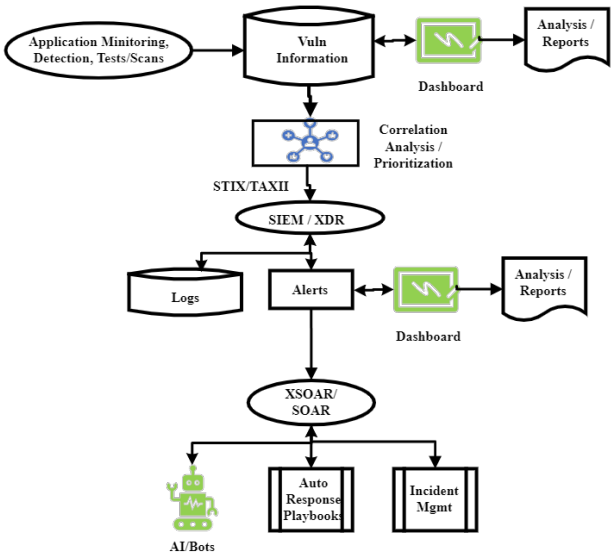


Figure 2: Conceptual Integrated of Vulnerability Data Platform and Incident Response Platform.

4. In Conclusion

This paper has proposed a data consolidation platform that will be foundational to build a comprehensive predictive system. We have also proposed how this data platform can be integrated with SIEM, SOAR, XSOAR, and XDR to establish a fully automated attack detection and response system. As mentioned in the paper, SIEM, SOAR, XSOAR, and XDR are more constructed to suit networks, infrastructure and devices, and sensors; not meant for application security vulnerability information as collected. So, this paper makes a special case that must be made for integration of application security information as part of threat intelligence and threat response.

Acknowledgements

This research has been supported and guided by Dr. Xiaolong Li of Indiana State University, USA, and Dr. John Pickard and Dr. Peng Li of East Carolina University, USA.

Dr. Xiaolong Li is a professor in the Department of Electronics and Computer Engineering Technology at Indiana State University. He received his PhD in Computer Engineering from the University of Cincinnati in 2006. His primary areas of research include modeling and performance analysis of MAC protocol, Internet of Things, Wireless Ad Hoc networks, and sensor networks.

Dr. John Pickard is a professor of Information and Cybersecurity Technology at East Carolina University, North Carolina, USA. He received his PhD in Technology Management from Indiana State University in 2014. His main research areas are internet protocols, convergence of information and operations technologies, and Internet of Things applications.

Dr. Peng Li received his Ph.D. in Electrical Engineering from the University of Connecticut. His professional certifications include CISSP, RHCE and VCP. Dr. Li is currently an Associate Professor at East Carolina University. He teaches undergraduate and graduate courses in programming, computer networks, information security, web services and virtualization technologies. His research interests include virtualization, cloud computing, cybersecurity, and integration of information technology in education.

References

Akamai, (2023) "Slipping Through the Security Gaps: The Rise of Application and API Attacks", *Akamai, https://www.akamai.com/blog/security/the-rise-of-application-and-api-attacks*

Carielli, S., DeMartine, A., Provost, A.C. and Dostie, P. (2021) “The Forrester Wave™: Software Composition Analysis, Q3 2021-The 10 Providers That Matter Most And How They Stack Up”, *Forrester, August, https://www.forrester.com/report/the-forrester-wave-software-composition-analysis-q3-2021/RES176091?ref_search=3502061_1674835391293&utm_source=PANTHEON_STRIPPED&utm_medium=email&utm_campaign=summit21na&utm_content=blog&categoryid=a89c0000000AKp1AAG%3Futm_source%3DPANTHEON_STRIPPED*

Carielli, S., DeMartine, A., Provost, A.C. and Dostie, P. (2022) "The Forrester Wave™: Web Application Firewalls, Q3 2022,The 12 Providers That Matter Most And How They Stack Up," *Forrester, September, https://www.forrester.com/report/the-forrester-wave-tm-web-application-firewalls-q3-2022/RES176396*

Checkmarx, (2022) "Correlation: The Application Security Testing Imperative in Modern Application Development", *Checkmarx,* https://checkmarx.com/resources/whitepapers/correlation-application-security-testing-imperative-in-modern-application-development/

Cinque, M., Cotroneo, D., and Pecchia, A. (2018) "Challenges and Directions in Security Information and Event Management (SIEM)", *2018 IEEE International Symposium on Software Reliability Engineering Workshops*

CrowdStrike, (2023) "2023 Global Threat Report," *CrowdStrike, https://www.crowdstrike.com/global-threat-report/*

Dynatrace, (2023) "The convergence of observability and security is critical to realizing DevSecOps potential," *Dynatrace CISO report, https://www.dynatrace.com/news/press-release/global-ciso-report-observability-security/*

EdgeScan, (2019) "Vulnerability Statistics Report," *pp. 4-17, Edgescan,* https://www.edgescan.com/wp-content/uploads/2019/02/edgescan-Vulnerability-Stats-Report-2019.pdf

FASTLY, (2022) "10 Key Capabilities of the Fastly Next-Gen WAF," *FASTLY, 2022, https://learn.fastly.com/security-10-key-capabilities-of-fastlys-next-gen-waf.html*

Gonzalez-Granadillo, G., Gonzalez-Zarzosa, S., and Diaz, R. (2021) "Security Information and Event Management (SIEM): Analysis, Trends, and Usage in Critical Infrastructures," *Sensors **2021**, 21, 4759. https://doi.org/10.3390/s21144759*

Johnson, J., McCarty, M., Richardson, B., Rieger, C., Cooley, R., Gentle, J.P., Rothwell, B., Phillips, T., Novak, B., Culler, M., Schwalm, K., and Wright, B. (2023) "Hardening Wind Energy Systems from Cyber Threats–Final Project Report", *SANDIA REPORT, SAND2023-12610, Printed February 2023*

Kasturi, S. (2020) "Post Implementation Evaluation of Coverage in Software Testing Using Monitoring Tools", *2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), Oct 2-4, 2020.*

Kasturi, S., Li, X., Pickard, J., and Li, P. (2023) "Understanding Statistical Correlation of Application Security Vulnerability Data from Detection and Monitoring Tools," *2023 33rd International Telecommunication Networks and Applications Conference*, *Melbourne, Australia, 2023, pp. 289-296, doi: 10.1109/ITNAC59571.2023.10368476.*

Mern, J., Hatch, K., Silva, R., Hickert, C., Sookoor, T., and Kochenderfer, M. J. (2022) "Autonomous Attack Mitigation for Industrial Control Systems", *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*

Muhammad, A.R., Sukarno, P., and Wardana, A.A. (2023) "Integrated Security Information and Event Management (SIEM) with Intrusion Detection System (IDS) for Live Analysis based on Machine Learning", *4th International Conference on Industry 4.0 and Smart Manufacturing, ScienceDirect, Procedia Computer Science 217 (2023) 1406–1415.*

Na, J. (2021)"Introducing Secure Application: True Runtime Application Self-Protection (RASP) for the Modern Application," CISCO App Dynamics,https://www.appdynamics.com/blog/product/application-security/

Nour, B., Pourzandi, M., and Debbabi, M. (2023) "A Survey on Threat Hunting in Enterprise Networks", *IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 25, NO. 4, FOURTH QUARTER 2023*

Olteanu, I. (2022) "Evaluating the response effectiveness of XDR technology in a scaled down environment", *Eindhoven University of Technology, Evaluating the response effectiveness of XDR technology in a scaled down environment — Eindhoven University of Technology research portal (tue.nl)*

Primeon, (2018) "Enterprise Applications: Wide Open to Attack in 2018," Primeon, https://www.primeon.com/whitepaper/primeon_wp2_r.pdf?1

Sevri, M. and Karacan, H. (2019) "Deep learning based web application security," in *Proc. of 2nd Int. Conf. on* Advanced Technologies, in Proc. Computer Engineering and Science (ICATCES), Antalya, Turkey, pp. 349-354, Apr. 2019.

SALT, (2023) "State of API Security Q1 2023, " *SALT LABS, https://content.salt.security/rs/352-UXR-417/images/SaltSecurity-Report-State_of_API_Security.pdf*

Signal Sciences, (2021) "Identifying Web Attack Indicators,". *https://info.signalsciences.com/identify-web-attack-indicators*

Velásquez, J.M.L., Monterrubio, S.M.M., Luis Enrique Sánchez Crespo, L.E.S., and Rosado, D.G. (2023) "Systematic review of SIEM technology: SIEM-SC birth", *International Journal of Information Security (2023) 22:691–711, https://doi.org/10.1007/s10207-022-00657-9*

Veracode, (2023) "State Of Software Security Vol. 11," *Veracode https://www.veracode.com/state-of-software-security-report*

Options for Signalling Cyber Deterrence Using Cyber Capabilities

Maria Keinonen[1] and Kimmo Halunen[1,2]
[1]National Defence University, Helsinki, Finland
[2] University of Oulu, Oulu, Finland

maria.keinonen@mil.fi
kimmo.halunen@oulu.fi

Abstract: The possibility of demonstrating power in cyberspace to create deterrence is a controversial topic. The desire of states to hide their true cyber capabilities leads to a reluctance to reveal their existing cyber power. However, the core idea of deterrence involves demonstrating power and signalling the will to use it so that the potential aggressor would find it less tempting to carry out aggression. Several studies attempt to solve the challenges related to cyber deterrence with a holistic approach, where deterrence in cyberspace is produced as part of a comprehensive deterrence using all instruments of state power, such as diplomatic, information, military, economic and legal capabilities. In turn, some studies argue that for credibility, cyber deterrence must include measures implemented specifically in cyberspace because cyberattacks can only be responded to in real-time with cyber capabilities. This paper argues that demonstrating cyber power is both necessary and profitable for the credibility of deterrence, although the nature of cyberspace and related technologies pose some limitations. This study examines the possibilities of demonstrating cyber power in concrete ways and aims to add a new perspective to academic debate. Cyber deterrence is investigated from the perspective of classical deterrence theory, including deterrence by denial and deterrence by punishment. By examining cyber deterrence literature using content analysis, deterrents that can be produced with cyber capabilities are defined, and examples of means to produce these effects are presented. According to the central observation of the study, a state can choose whether to demonstrate cyber power by revealing the victories achieved in real-life cyber battles, by demonstrating force in another state's cyberterritory or by disclosing selected capabilities in separate simulations without a real-life connection.

Keywords: Cyber deterrence, Cyber power, Cyber deterrence signalling, Demonstration of cyber power

1. Introduction

Cyber deterrence can be understood in at least three different ways. First, it can be understood as using the state's cyber power as a deterrent against a military attack. The second conception is using military resources as a deterrent against a cyberattack. Third, it may mean using the state's cyber capabilities as a deterrent against a cyberattack. (Soesanto & Smeets, 2021) This paper investigates the third option from the perspective of deterrence signalling.

Cyber deterrence research is focused on controlling real-life cyber events and demonstrating power with cyber operations (Bendiek & Metzger, 2015). Only a few studies focus on the enablers of cyber operations, such as governance, training and decision-making (Huskaj, 2019). Also, displaying cyber power that does not involve real-life cyber operations has received less attention. This research explores the possibilities of demonstrating cyber power in real-life events and other ways that do not require the disclosure of the state's critical cyber capabilities.

Due to the unique characteristics of cyberspace, there are still many unsolved challenges associated with cyber deterrence. Some of these challenges could be resolved with already existing implementation options if consciously combined with deterrence signalling. These options are investigated in this article. The research question is: *"What kind of options can be utilised for cyber deterrence signalling?"* To answer the question, the study analyses cyber deterrence literature and forms an understanding of the possibilities and limitations of cyber deterrence signalling. This analysis sets the conditions for solutions sought from real-life cyber events as examples of how cyber deterrence can be signalled.

2. Theoretical Background

At the core of the classical deterrence theory is the idea of force, which the state is ready to use against aggressions by other states. The will to use this force is communicated preventively to decrease the threat actor's desire to behave aggressively. With the changing power relations in world politics and the development of cyberspace, the need for deterrence has remained. However, the concept has changed from absolute nuclear deterrence to a broad spectrum of deterrents and flexible escalation control. (Freedman, 2021)

A fundamental dichotomy in classical deterrence literature appears in subsequent deterrence studies in one form or another. According to these studies, deterrence can be divided into deterrence by denial and deterrence by punishment. The goal of deterrence by denial is to convince the threat actor that aggression will

not achieve the desired effects and goals. The main message of deterrence by punishment is that at least an equal response follows that aggression. The classical deterrence theory assumes that deterrence is implemented by states that signal their intentions to each other. Deterrence will then fail if aggressions are carried out. (Mazarr, 2021)

The principles of classical deterrence theory do not apply to cyberspace as such (Taddeo, 2018), and it is no longer a dichotomous concept where it either succeeds or fails depending on the aggressions that have taken place (Chen, 2023). Cyberspace will not achieve complete immunity from attacks because it offers more opportunities for attack than opportunities for protection. (Tor, 2017) However, this is not a valid reason not to engage in deterrence signalling related to the protection of cyberspace sovereignty.

Part of deterrence signalling is a demonstration of force. Warfighting is the state's most unambiguous opportunity to demonstrate its military power. Since deterrence aims to prevent aggression, force must be demonstrated before tensions escalate into an armed conflict. In peacetime, states can demonstrate their military power with military exercises and deliberately promote experiments and exhibitions. These means can also be used for abilities that are more difficult to demonstrate, such as cyber capabilities. (Montgomery, 2020)

The classical deterrence theory requires the ability to respond to aggression. That, in turn, requires the ability to identify the attacker reliably. In cyberspace, attribution can be challenging if techniques have been used that make it possible to hide the origin of the attack. Although the development of technology has enabled more effective attribution, the state rarely wants to report the attack to hide its actual cyber capabilities. (Lonergan & Lonergan, 2023; Chen, 2018; Taddeo, 2018; Wanic & Rowe, 2018; Lee, 2015) On the other hand, developing technology to enable attribution is profitable for the state because successful attribution also increases the credibility of cyber deterrence (Navicky & Tkach, 2023; Baliga et al., 2018; Chen, 2018). Therefore, cyber deterrence signalling via real-life actions is problematic because states want to hide their cyber capabilities. At the same time, deterrence is based on publicised facts of capability and will to use it.

Another challenge of cyber deterrence is credibility, the lack of which renders signalling useless. In contrast to nuclear weapons and conventional military power, cyber capabilities represent soft power that can hardly directly achieve massive destructive power, as required by deterrence by punishment. (Lonergan & Lonergan, 2023; Chen, 2018, Taddeo, 2018; Lee, 2015) For this reason, cyber deterrence signalling must be based on factors other than brute force.

For the challenges presented, the academic discussion on cyber deterrence still debates different implementation options (Soesanto & Smeets, 2021). Cyber persistence theory offers a novel solution to protecting a state's sovereignty in cyberspace by continuously seeking initiatives to create and maintain favourable conditions of security in cyberspace. This approach is based on principles of exploitation rather than coercion (Fischerkeller et al., 2022). Another approach is cyber diplomacy, which enhances security and stability in cyberspace via cooperation, confidence-building and the establishment of international norms. (van der Meer, 2015) Although these approaches to state cyber security differ in principle from cyber deterrence, they offer tools that can also be utilised from the perspective of cyber deterrence, for example, the ideas of initiative, continuity and cooperation.

3. Methodology and Results

This research focuses on deterrence signalling using cyber capabilities to defend the state's sovereignty in cyberspace. It should be noted, however, that deterrence signalling includes communication about the readiness to use all the instruments of the state's power to defend its sovereignty, such as diplomatic, information, military, economic and legal measures (Sweijs & Zilincik, 2021). Together with these, the demonstration of cyber power completes the entirety of the state's deterrence signalling.

The study's theoretical basis was formed by content analysis (Puusa & Juuti, 2020). Scientific research on cyber deterrence was selected as the source material and searched in abstract and citation databases Scopus and Google Scholar. The search phrases included "cyber deterrence" and "deterrence in cyberspace". These phrases were searched from the title, keywords and abstract. The period was limited between 2013 and 2023 to form an understanding of earlier research and current views on the subject. The initial search produced 209 hits.

The screening process narrowed the focus to articles about deterrence by denial, deterrence by punishment or deterrence signalling. After screening, twenty scientific studies that focused on the possibilities and limitations

of cyber deterrence and its implementation possibilities were selected as the material to be analysed. After the first analysis round, themes were formed, and the second analysis round was based on these themes. The themes were: "features of cyber deterrence", "options for demonstrating cyber power", and "deterrence options". Categories were formed under the selected themes.

Six features of cyber deterrence were defined based on the analysed material. These features are presented in Table 1. The "x" marks the information found in each article. These features apply to demonstrating both defensive and offensive cyber capabilities.

Table 1: Results of the content analysis: part 1.

	Features of cyber deterrence					
Factor **Writer**	**Continuous actions**	**Co-operation and confidence building**	**The level of secrecy is a choice**	**Attribution is a key capability**	**Need for escalation control**	**Deterrence is not absolute**
Borghard & Lonergan, 2023	x		x			x
Chen, 2023			x	x	x	
Lonergan & Lonergan, 2023		x	x			
Navicky & Tkach, 2023	x		x	x	x	
Pedersen, 2023			x	x	x	
Brown & Fazal, 2021			x			
Kostyuk, 2021			x		x	
Klimburg, 2020		x	x		x	
Montgomery, 2020			x		x	
Baram & Sommer, 2019		x	x		x	
Fischer, 2019			x	x	x	x
Baliga et al, 2018			x	x		
Chen, 2018			x	x	x	
Taddeo, 2018			x	x	x	
Wanic & Rowe, 2018		x	x	x		x
Carson & Yarhi-Milo, 2017			x		x	
Edwards et al, 2017			x	x	x	
Tor, 2017	x					x
Lee, 2015			x		x	
Lindsay, 2015			x	x	x	

Continuous cyber operations signal to the threat actor about the state's maturity to protect itself from cyberattacks, making it difficult for the threat actor to carry out the aggression. Such continuity can reduce the threat actor's desire to attack. (Borghard & Lonergan, 2023) Such deterrence signalling is based on long-term activity and demonstrating determination and force in cyberspace.

International cooperation in developing cybersecurity-related laws and norms and responding to cyberattacks together signals to the threat actor the will to solve cyber threats as an international front. (Klimburg, 2020; Wanic & Rowe, 2018) Sharing information about one's cyber capabilities can instil confidence in allies and neutral parties and increase stability with a potential threat actor. (Lonergan & Lonergan, 2023; Klimburg, 2020) Signalling these actions specifically as part of deterrence could enhance credibility.

It is not advantageous for the state to reveal too much of its cyber capabilities or activities in cyberspace, so a certain amount of secrecy is needed, for example, to protect the gained foothold in opponent systems and reconnaissance (Navicky & Tkach, 2023; Baram & Sommer, 2019). A state should not reveal anything it might

want to take advantage of later (Lonergan & Lonergan, 2023). The most risk-free option is revealing facts without connection to the state's actual cyber capabilities.

Attribution strengthens if the state continuously improves its abilities for detection and identification (Baliga et al., 2018). If the state has built a credible attribution capability and publicly informs about the identified attackers, it may be easier to justify counterattacks in the eyes of the international communities (Wanic & Rowe, 2018). Correspondingly, repeated failures to identify an attacker inevitably also lead to a weakening of deterrence (Navicky & Tkach, 2023). From a signalling point of view, attribution is a crucial capability in cyberspace.

Although cyber capabilities are considered soft measures, there could be a risk of escalation if a cyberattack is answered with a counterattack. Attacks are also accompanied by uncertainty in controlling the spread of the effects of cyberattacks. (Taddeo, 2018; Lee, 2015) Therefore, there is a need for escalation control when using offensive cyber capabilities.

The state must accept that deterrence against cyberattacks is not absolute. Even so, it is worth making the cyberattack as complex and resource-consuming as possible because it can have a deterrent effect. (Fischer, 2019; Tor, 2017) Despite this, the state must signal its determination to defend itself against cyberattacks.

During the analysis, it was found that several cyber deterrence studies have focused on real-life cyber activity. This activity can be divided into countermeasures against attacks on the state and its offensive cyber operations against another state or a display of cyber power in which the state shows its power through a third party. These categories were named real-life events and proxy events in Table 2.

A real-life event can be offensive or defensive. A fundamental choice is the level of secrecy, which is also related to escalation control. For example, while demonstrating the attribution ability, one can decide to share all the related information or only about the events and their prevention, leaving the attacker's identity undisclosed (Baram & Sommer, 2019). The latter choice can help manage the risk of escalation, primarily if the communication focuses on established evidence and events, leaving sensitive issues unaddressed (Lindsay, 2015). Accordingly, when planning and executing offensive cyber operations, the state must assess the chances of getting caught and the resulting consequences (Edwards et al., 2017) and, if caught, choose whether to deny or admit guilt or even not comment on the matter at all (Brown & Fazal, 2021). Successful cyberattacks signal the state's ability and will to use force in cyberspace but might cause escalation and inevitably reveal the intelligence's footholds in opponent systems and techniques used for the attack.

Proxy events can be, for example, the sharing of information that targets a third party. When a state detects a cyberattack, it can share related findings with the target state or participate in combating and countering the attack (Klimburg, 2020; Wanic & Rowe, 2018). Through a third party, a commercial operator or another authority, it is also possible to reveal information the state does not want to associate with, for example, military intelligence capability.

The analysis found that cyber power can also be demonstrated with simulated events. These can be used to demonstrate cyber power without risking the exposure of the state's true cyber capabilities. The aim is to demonstrate the state's maturity as a cyberspace actor, including signalling expertise, technological capabilities and cooperation. Simulated events can be, for example, national and international cyber exercises, competitions and public demonstrations, which are communicated openly (Montgomery, 2020; Wanic & Rowe, 2018).

Table 2 presents the prevalence of real-life, proxy and simulated events in demonstrating cyber power and the author's point of view on the deterrence of denial or punishment in the analysed material.

According to the analysis, the basic principle for signalling deterrence by denial in cyberspace is demonstrating the ability to withstand and repel cyberattacks. Correspondingly, deterrence by punishment is signalled by demonstrating the will and ability to attack. The purpose of demonstrating cyber power is not always to inflict damage to the opponent but also to demonstrate the state's maturity to act in cyberspace. Table 3 summarises examples related to the possibilities of demonstrating cyber power for both types of deterrence. The selected examples are existing state activities and have been investigated from the perspective of real-life, proxy and simulated events.

Table 2: Results of the content analysis, part 2

	Options for demonstrating cyber power			Deterrence options		
Factor **Writer**	Real-life events	**Proxy events**	**Simulated events**	**Denial with counter-operations or retaliation**	**Denial with resilience and defence**	**Punishment with offensive cyber operations**
Borghard & Lonergan, 2023	x			x		
Chen, 2023				x		x
Lonergan & Lonergan, 2023	x					
Navicky & Tkach, 2023	x				x	
Pedersen, 2023	x					x
Brown & Fazal, 2021	x					
Kostyuk, 2021		x			x	x
Klimburg, 2020		x				
Montgomery, 2020			x			
Baram & Sommer, 2019	x					
Fischer, 2019	x			x	x	
Baliga et al, 2018				x		
Chen, 2018					x	x
Taddeo, 2018	x				x	x
Wanic & Rowe, 2018	x	x	x	x	x	x
Carson & Yarhi-Milo, 2017	x					
Edwards et al, 2017	x					
Tor, 2017	x			x		x
Lee, 2015					x	
Lindsay, 2015	x				x	x

The purpose of the simulated event is to demonstrate the state's maturity to protect its sovereignty in cyberspace via defensive and offensive means without disclosing the actual cyber capabilities. This can be achieved via cyber exercises, competitions and other simulations in training environments such as cyber ranges. They reflect either denial or punishment, depending on the type of the event.

A cyber range includes a realistic network environment with realistic scenarios, enabling various roles and teaming, feedback for learning purposes and monitoring the trainee's actions during the exercise. (Muhammad et al., 2020) A state can utilise cyber ranges for signalling deterrence by denial or punishment by publicly disclosing information about their use.

With cyber exercises, states can demonstrate their expertise, technological capabilities and ability to cooperate with other states. Locked Shields is an international cyber exercise conducted by the NATO Cooperative Cyber Defence Centre of Excellence (CCDCOE); it is defensive in nature and allows states to conduct strategic signalling. (Smeets, 2022) From the deterrence by punishment point of view, the CCDCOE's Crossed Swords exercise incorporates red teaming techniques, tools, tactics and procedures (Blumbergs et al., 2019), which can be used for demonstrating offensive capabilities. Also, hacking competitions like the Pwn2Own (Portnoy, 2010) allow governments to demonstrate offensive cyber capabilities. It is profitable for states to signal their participation and success in such competitions, even if their more detailed content cannot be published.

From the point of view of deterrence by denial, a proxy event can occur, for example, by sharing information about detected attacks against another state or aiding another state to identify and counter a cyberattack. The war in Ukraine has shown the possibility of assisting another country to defend itself through collaboration

with Western nations and technology companies (Willet, 2022). This is a valuable opportunity for the supporting country to signal its ability to execute cyber operations on the territory of another state, and vice versa, an opportunity for the defender to demonstrate its abilities to act as a host country in a multinational cyber defence operation.

Signalling the states' ability to recruit patriotic hackers for offensive cyber operations benefits deterrence by punishment as a proxy event. For example, Russia published attacking guidelines against Georgia's government websites in 2008 and successfully recruited patriotic hackers for the task (Mareš & Netolická, 2020). Another example is the voluntary organisation "IT Army of Ukraine", established by the Ukrainian government, which allows cyber experts from Ukraine and other countries to execute cyber operations (Soesanto, 2023). On the other perspective, when recruiting civilians, the level of control can be challenging to establish, and some of the effects can be detrimental to the overall operation. This could also decrease the credibility of deterrence signalling.

Isolating a state's internal network from the global Internet is an extreme example of the practical implementation of real-life deterrence by denial. Such an example is Russia's RuNet. According to the plans and outspoken rhetoric, RuNet is intended to be an autonomous network that can be taken off the Internet without affecting RuNet's internal networking (Kukkola, 2020; Kukkola, 2018). Western media have commented that Russia has successfully tested this capability (Mellor, 2022; Wakefield, 2019). Even though no evidence of this has been publicly presented, talking about it can be interpreted as Russia's cyber deterrence signalling.

An alternative strategy to deterrence by denial is investing in state resilience and cyber diplomacy. For example, the European Union actively promotes the development of the international regulatory framework and projects its normative power (Miadzvetskaya & Wessel, 2022). Resilience is developed through, among other things, disruptive technologies in building a defensive shield for the EU member states. (Osula, 2022) Communicating about resilience and cyber diplomacy also strengthens deterrence signalling.

From deterrence by punishment perspective, there are examples of cyberattacks affecting cyber and physical environments. Stuxnet, released in 2010, was used to destroy critical infrastructure related to nuclear enrichment in Iran. It was in Israel's and the United States' interests to halt Iran's nuclear program, so they executed a cyber campaign code-named "Olympic Games". (Lindsay, 2013) As a result, despite the states' motives, Iran is not yet a nuclear weapon state. Therefore, the Stuxnet may have played a role in deterrence by punishment, convincing Iran that more severe consequences might occur if the nuclear program were to proceed.

Table 3: Options for signalling cyber deterrence using cyber capabilities

Deterrence strategy / Execution option	Deterrence by Denial	Deterrence by Punishment
Simulated event	*Demonstrations of expertise, technological readiness and cooperation via cyber exercises and competitions.* Examples: Cyber Ranges, Locked Shields.	*Demonstrations of expertise, technological readiness and cooperation via cyber exercises and competitions* Examples: Crossed Swords, Cyber ranges, Pwn2Own.
Proxy event	*Information sharing of detected attacks against another state. Aiding another state to identify and counter a cyberattack. The ability to receive support.* Examples: Ukraine cyber defence collaboration.	*Information sharing of potential aggressors' networks. Offensive cyber operation with another state or targeting an enemy of another state.* Examples: IT Army of Ukraine, patriotic hackers.
Real-life event	*Publicly shared information.* Examples: RuNet, EU cyber diplomacy and resilience.	*Publicly shared information.* Examples: Stuxnet, NotPetya

NotPetya, released in 2017, was originated by Russian military intelligence targeting the Ukrainian economy by encrypting and paralysing the computer networks of Ukrainian banks, firms, and government. (Crosignani et

al., 2021) Effective punishment in cyberspace generally requires technical capabilities and an understanding of the adversary's target systems. This ability could considerably threaten a potential aggressor alongside other deterrence signalling.

4. Conclusions

This article presents three options for demonstrating cyber power: simulated, proxy and real-life events. These options were examined from the point of view of deterrence by denial and deterrence by punishment, looking for implementation examples for each option.

Simulated events, for example, cyber exercises and competitions, can demonstrate the state's ability to protect its sovereignty in cyberspace without revealing the state's actual cyber capabilities. A lack of direct connection to real-life activities can be a weak deterrence signal. However, continuous signalling could create a reputation of states' cyber maturity and benefit deterrence strategies.

Proxy events can be, for example, information sharing of detected attacks against another state and aiding another state to identify and counter a cyberattack or targeting a mutual enemy. This allows the deterring state to demonstrate its ability to collect information and operate on other states' cyber territory and the defending state to demonstrate its ability to receive assistance.

Real-life events are often more straightforward to signal from the perspective of deterrence by denial because related cyber activities can often be signalled without revealing technological details and concentrating on the achieved effects. Demonstrating offensive capabilities can be more challenging because the cyber attacker often wants to hide the attack's origin. Since the level of secrecy is often a choice, the state must consider the cost-benefit calculus when using real-life events for deterrence signalling.

In conclusion, existing options exist for demonstrating cyber power through simulated, proxy and real-life events. States need to recognise these options and use them systematically for deterrence signalling. From a signalling point of view, the presented options for demonstrating cyber power differ in signal clarity and credibility. Despite this, it is worthwhile to demonstrate cyber power in as many ways as possible because these signals can create a credible image of the state's cyber deterrence.

References

Baliga, S., Bueno De Mesquita, E. & Wolitzky, A. (2020). Deterrence with Imperfect Attribution. *American Political Science Review*, 114(4), 1155-1178. doi:10.1017/S0003055420000362.

Baram, G. and Sommer, U. (2019). Covert or not Covert: National Strategies During Cyber Conflict. *2019 11th International Conference on Cyber Conflict (CyCon),* Tallinn, Estonia, 2019, pp. 1-16, doi: 10.23919/CYCON.2019.8756682.

Bendiek, A., & Metzger, T. (2015). *Deterrence theory in the cyber-century. Lessons from a state-of the-art literature review*. Working Paper, Research Division EU/Europe, Stiftung Wissenschaft und Politik, German Institute for International and Security Affairs.

Blumbergs, B., Ottis, R. & Vaarandi, R. (2019). Crossed Swords: A Cyber Red Team Oriented Technical Exercise. *18th European Conference on Cyber Warfare and Security*, Volume 2019-July, Pages 37 – 44.

Borghard, E. & Lonergan, S. (2023). Deterrence by denial in cyberspace, *Journal of Strategic Studies*, 46:3, 534-569, DOI: 10.1080/01402390.2021.1944856.

Brown, J., & Fazal, T. (2021). #SorryNotSorry: Why states neither confirm nor deny responsibility for cyber operations. *European Journal of International Security*, 6(4), 401-417. doi:10.1017/eis.2021.18.

Carson, A. & Yarhi-Milo, K. (2017). Covert Communication: The Intelligibility and Credibility of Signaling in Secret. *Security Studies*, 26:1, 124-156, DOI: 10.1080/09636412.2017.1243921.

Chen, J. (2023). Deterrence in Cyberspace: An Essential Component in Integrated Deterrence. In Billingsley, J. (ed), *Integrated Deterrence and Cyberspace, Selected Essays Exploring the Role of Cyber Operations in the Pursuit of National Interest*. National Defense University Press, Washington, D.C.

Chen, J. (2018). Does Conventional Deterrence Work in the Cyber Domain? *ECCWS 2018 17th European Conference on Cyber Warfare and Security V2*, Norway.

Crosignani, M., Macchiavelli, M. & Silva, A. (2021). *Pirates without Borders: The Propagation of Cyberattacks through Firms' Supply Chains*. Federal Reserve Bank of New York Staff Reports, no. 937.

Edwards, B., Furnas, A., Forrest, S. & Axelrod, R. (2017). Strategic aspects of cyberattack, attribution, and blame. *Proceedings of the National Academy of Sciences*, vol. 114, no. 11, pp. 2825–2830, 10.1073/pnas.1700442114.

Fischer, M. (2019). The Concept of Deterrence and Its Applicability in the Cyber Domain. *Connections QJ* 18, no. 1-2 (2019): 69-92. https://doi.org/10.11610/Connections.18.1-2.05.

Fischerkeller, M., Goldman, E., Harknett, R. (2022). *Cyber Persistence Theory: Redefining National Security in Cyberspace*. Oxford University Press, 2022. https://doi.org/10.1093/oso/9780197638255.001.0001.

Freedman, L. (2021). Introduction—The Evolution of Deterrence Strategy and Research. In: Osinga, F., Sweijs, T. (eds) *NL ARMS Netherlands Annual Review of Military Studies 2020*. NL ARMS. T.M.C. Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-419-8_8.

Huskaj, G. (2019). The Current State of Research in Offensive Cyberspace Operations. *Proceedings of the 18th European Conference on Cyber Warfare and Security*, 2019, pp. 660–667.

Klimburg, A. (2020). Mixed Signals: A Flawed Approach to Cyber Deterrence. *Survival*, 62:1, 107-130, DOI: 10.1080/00396338.2020.1715071.

Kostyuk, N. (2021). Deterrence in the Cyber Realm: Public versus Private Cyber Capacity. *International Studies Quarterly*, Volume 65, Issue 4, December 2021, Pages 1151–1162, https://doi.org/10.1093/isq/sqab039 .

Kukkola, J. (2018). Civilian and military information infrastructure and the control of the Russian segment of Internet. *2018 International Conference on Military Communications and Information Systems (ICMCIS),* Warsaw, Poland, 2018, pp. 1-8, doi: 10.1109/ICMCIS.2018.8398700.

Kukkola, J. (2020). *Digital Soviet Union: the Russian national segment of the Internet as a closed national network shaped by strategic cultural ideas*. National Defence University, Series 1: Research Publications No. 40, Doctoral Dissertation. Tampere: PunaMusta.

Lee, W.H., (2015). The Challenges of Cyber Deterrence. *Pointer, The Journal of the Singapore Armed Forces*. Vol.41 NO.1. ISSN 2017-3956.

Lindsay, J. (2015). Tipping the scales: the attribution problem and the feasibility of deterrence against cyberattack. *Journal of Cybersecurity*, 1(1), 2015, 53–67. doi: 10.1093/cybsec/tyv003.

Lindsay, J. (2013). Stuxnet and the Limits of Cyber Warfare. *Security Studies*, 22:365–404, 2013. doi: 10.1080/09636412.2013.816122.

Lonergan, E., & Lonergan, S. (2023). *Escalation Dynamics in Cyberspace*, Oxford University Press, Incorporated. ProQuest Ebook Central, https://ebookcentral.proquest.com/lib/ndul/detail.action?docID=7214000 .

Mareš, M. & Netolická, V. (2020). Georgia 2008: Conflict Dynamics in the Cyber Domain, *Strategic Analysis*, 44:3, 224-240, DOI: 10.1080/09700161.2020.1778278.

Mazarr, M. (2021). Understanding Deterrence. In: Osinga, F., Sweijs, T. (eds) *NL ARMS Netherlands Annual Review of Military Studies 2020*. NL ARMS. T.M.C. Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-419-8_8.

Mellor, S. (2022). *Experts say Russia's war on Ukraine is accelerating the 'splinternet.' But what is the splinternet?* Fortune. https://fortune.com/2022/03/22/russia-war-ukraine-great-firewall-splinternet-internet/ .

Miadzvetskaya, Y. & Wessel, R. (2022). The Externalisation of the EU's Cybersecurity Regime: The Cyber Diplomacy Toolbox. *7(1) European Papers 2022*, University of Groningen Faculty of Law Research Paper No. 27/2022. https://ssrn.com/abstract=4199627 .

Montgomery, E. (2020). Signals of strength: Capability demonstrations and perceptions of military power, *Journal of Strategic Studies*, 43:2, 309-330. doi: 10.1080/01402390.2019.1626724.

Muhammad, M., Basel K. & Vasileios G. (2020). Cyber ranges and security testbeds: Scenarios, functions, tools and architecture, *Computers & Security*, Volume 88/ 2020. https://doi.org/10.1016/j.cose.2019.101636.

Navicky, M. & Tkach, B. (2023). Cross-Domain Cyber Incidents and State Responses. In Billingsley, J. (ed), *Integrated Deterrence and Cyberspace, Selected Essays Exploring the Role of Cyber Operations in the Pursuit of National Interest*. National Defense University Press, Washington, D.C.

Osula, AM. (2022). Building Cyber Resilience: The Defensive Shield for the EU. In: Boulet, G., Reiterer, M., Pardo, R.P. (eds) *Cybersecurity Policy in the EU and South Korea from Consultation to Action. New Security Challenges*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-08384-6_9.

Portnoy, A. (2010). Pwn2Own wrap up and analysis. *Network Security*, Volume 2010, Issue 4,2010, Pages 4-5, https://doi.org/10.1016/S1353-4858(10)70043-X.

Puusa, A. & Juuti, P. (2020). *Laadullisen tutkimuksen näkökulmat ja menetelmät*. Helsinki: Gaudeamus.

Smeets, M. (2022). The Role of Military Cyber Exercises: A Case Study of Locked Shields, *2022 14th International Conference on Cyber Conflict: Keep Moving! (CyCon),* Tallinn, Estonia, 2022, pp. 9-25, doi: 10.23919/CyCon55549.2022.9811018.

Soesanto, S. & Smeets, M. (2021). Cyber Deterrence: The Past, Present, and Future. In: Osinga, F., Sweijs, T. (eds) *NL ARMS Netherlands Annual Review of Military Studies 2020.* NL ARMS. T.M.C. Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-419-8_20.

Soesanto, S. (2023). Ukraine's IT Army, *Survival*, 65:3, 93-106, doi: 10.1080/00396338.2023.2218701.

Sweijs, T. & Zilincik, S. (2021). The Essence of Cross-Domain Deterrence. In: Osinga, F., Sweijs, T. (eds) *NL ARMS Netherlands Annual Review of Military Studies 2020.* NL ARMS. T.M.C. Asser Press, The Hague. https://doi.org/10.1007/978-94-6265-419-8_8.

Taddeo, M. (2018). The Limits of Deterrence Theory in Cyberspace. *Philos. Technol*. 31, 339–355 (2018). https://doi.org/10.1007/s13347-017-0290-2.

Tor, U. (2017). 'Cumulative Deterrence' as a New Paradigm for Cyber Deterrence. In *Journal of Strategic Studies*, 40:1–2, (s. 92–117). https://doi.org/10.1080/01402390.2015.1115975.

Van der Meer, S. (2015). Enhancing International Cyber Security: A Key Role for Diplomacy. *Security and Human Rights.* 26. 193-205. https://doi.org/10.1163/18750230-02602004 .

Virtanen, T. & Simola, P. (2022). Layer 8 Tarpits: Overwhelming Malicious Actors with Distracting Information. In *Proceedings of the 21st European Conference on Cyber Warfare and Security.*

Wakefield, J. (2019). *Russia 'successfully tests' its unplugged internet*. BBC news. https://www.bbc.com/news/technology-50902496 .

Wanic, E. & Rowe, N. (2018). Assessing Deterrence Optinos for Cyber Weapons. *2018 International Conference on Computational Science and Computational Intelligence (CSC*I), Las Vegas, NV, USA, 2018, pp. 13-18, doi: 10.1109/CSCI46756.2018.00011.

Willett, M. (2022). The Cyber Dimension of the Russia–Ukraine War, *Survival*, 64:5, 7-26, DOI: 10.1080/00396338.2022.2126193.

A Federated Distributed Digital Forensic Readiness Model for the Cloud

Renico Koen and Hein Venter
University of Pretoria, South Africa

Renico.koen@gmail.com
hein.venter@up.ac.za

Abstract: Digital forensics in modern, cloud-based, microservice-based applications are complicated by multiple layers of abstraction, thereby making it difficult to accurately capture and correlate events that occur across these layers due to filtering caused by abstraction. The complexities linked to each layer of abstraction are primarily invisible to subsequent layers. Similarly, software services are often composed of one or more services provided by various service providers across the globe. Investigators are often faced with situations where breaches span over multiple service provider boundaries where not all digital forensic readiness evidence artefacts are captured by the service provider's forensic readiness processes. Instead, digital evidence artefacts are scattered across multiple service provider domains. This paper presents a novel, federated distributed digital forensic readiness model suitable for use in software-as-service, platform-as-service and infrastructure-as-service provider scenarios. The proposed model enables a service provider to capture and inspect forensic readiness artefacts in environments with various layers of abstraction. More importantly, the model also offers a way to share and access forensic readiness artefacts in a forensically sound manner to ultimately ensure that investigators can obtain a clear view of digital forensic events as they occur between amalgamated services provided by one or more separate service providers.

Keywords: Digital forensic readiness, Digital forensics, Cloud computing, Information silos

1. Introduction

Cloud computing is often called one of the most transformative technologies in recent history due to the way in which services can be composed and consumed (Ruan, et al., 2011). Cloud computing describes highly scalable, on-demand computing resources offered by service providers on a pay-as-you-go basis.

Cloud computing has been mentioned to reduce operational costs and improve operational efficiencies (Armbrust, et al., 2010). With the traditional client/server model of computing, i.e. without a cloud environment, server infrastructure must be purchased and maintained when hosting applications. The maintenance of these servers requires constant action from skilled team members. When calculating the total cost of ownership of traditional hosting models, the argument can be made that many hidden costs are typically present that impact the total cost of ownership of a provisioned service.

Cloud computing is convenient because many of the functions traditionally required for a client/server model are already included as part of the cloud computing package. Consumers end up paying only for services that they consume (Shetty, et al., 2014). Some cloud computing benefits include resilience, reliability, and on-demand access (Shetty, et al., 2014).

However, cloud computing does pose various challenges in terms of forensic readiness. The purpose of this paper is to highlight current digital forensic readiness challenges in cloud environments. A federated model is proposed to help address some of the challenges highlighted.

The rest of this paper is structured as follows: background information is discussed in section 3. Related work is discussed in section 4 and cloud computing challenges are highlighted in section 5. A federated digital forensics model is proposed in section 6. A discussion of the proposal is presented in section 7 while future work is discussed in section 8. The chapter is concluded in section 9.

2. Background

There are three recognised cloud service environments, namely infrastructure-as-a-service (IaaS), Platform-as-a-service (PaaS) and Software-as-a-service (SaaS) (Mell & Grance, 2011) and (Liu, et al., 2011) . IaaS is concerned with the provisioning of virtualised infrastructure, such as network facilities, storage facilities and virtual machines (Alenezi, et al., 2017). The infrastructure provided by IaaS is relatively primitive but can be used to build more complex offerings, such as PaaS and SaaS. Villegas et al. (2012) refer to IaaS as the infrastructure resources used by the rest of the service stack to provide services. The customer does not manage or control the underlying hardware but can control the operating system and deployed applications (Mell & Grance, 2011).

The result is that the customer can focus on providing services that are within their fields of expertise while consuming infrastructure services provided by infrastructure specialists.

PaaS is concerned with providing a set of hosted libraries and tools (Alenezi, et al., 2017). The provided platforms allow service companies to create service offerings based on the provided platforms. The provided platforms are hosted on IaaS (typically included as part of the PaaS offering). Widely known examples of PaaS include Microsoft Azure and the Google App Engine (Villegas, et al., 2012) and AWS. Facilities offered by PaaS are limited to the components that have been assembled or developed by the platform provider (Weir, et al., 2018). The consumer does not manage the underlying hardware or operating system environment but can deploy applications and change application hosting settings (Mell & Grance, 2011). As a result, the customer is shielded from most of the technicalities related to hardware and underlying operating system management and securing of the underlying resources.

SaaS can be described as hosted applications (Alenezi, et al., 2017). These applications are typically provided via the user browser or API (Alenezi, et al., 2017). The SaaS offerings typically include PaaS (and, as a result, IaaS) components as part of a managed service. SaaS offers various benefits, such as shorter application implementation time and the owners' elimination of software version management (as the service provider manages version management.

Cloud computing helps organisations deliver services or access online services that are often amalgamations of services provided by various vendors. However, the investigation of security events in service delivery environments is often complicated because multiple service providers may be present, which can potentially be located all over the world and are governed by different legal frameworks.

The next section of this paper discusses current challenges in cloud forensics.

3. Related Work

Current digital forensics challenges can be classified into five categories, namely complexity of low-level data acquired, diversity of data sources, consistency and correlation problem caused by current tooling that attempts to find fragments of evidence and nothing more, and volume and unified timelining issues caused by various evidence sources acquired using non-consistent time zones and time sources (Lillis, et al., 2016). The challenges were discussed by Lilllis et al. (2016), who also highlighted that the Internet of Things would likely worsen existing digital forensics data processing backlogs due to increases in volumes of data. In addition to these categories, there are significant technological challenges to contend with. Technological challenges include the rapidly changing technological landscape, encryption, large volumes of data and the complexity of digital systems and networks (Alenezi, 2023).

Miller et al. (2014) proposed a cloud-based distributed processing platform called Forensicloud in an attempt to address the increasing volume and diversity of data in forensic examinations.

Furthermore, Duijn & Sloot (2015) explore the use of big data and analytics to determine how organised crime groups operate and adapt over time. Analysts may be able to uncover patterns that are hidden or not obvious by analysing data from a diverse set of sources. However, the quality of data sources plays a role in the output produced.

Monteiro et al. (2023) discussed challenges related to forensics in a microservice environment. The author emphasised the importance of forensic-ready microservices and proposed a framework that incorporates game theory to achieve such a goal.

The lack of standardisation, the complexity of cloud environments and the dynamic nature of cloud environments are typical problems in cloud environments (Alenezi, 2023). Hence, cloud service providers must implement their services in a way that ensures that investigators can gather evidence in a forensically sound manner (Simou, et al., 2022). As a result, an organisation's forensic readiness levels should be monitored to ensure readiness should an incident occur.

Shanmugasundaram et al. (2003) proposed a distributed network logging mechanism named ForNet, to assist digital forensics over a wide area network. A component named SynAps acts as an agent that summarises network events. These events are then sent to a server (called Forensic Server) for analysis.

Federici (2013) proposed an architecture called AlmaNebula with the purpose of providing forensics as a service for cloud computing by utilising commodity machines for processing large amounts of data.

Sibiya et al. (2013) discussed security challenges related to cloud computing and digital forensics. A forensic readiness model was introduced that utilises forensics services hosted in the cloud to minimise the amount of time needed to perform forensic investigations. The model consists of various components, such as application forensics, memory forensics, network forensics and computer forensics.

Kebande & Venter (2018) proposed cloud forensic readiness as a service model for collecting and preservice potential digital evidence. Agent-based sets are used to collect information. Collected evidence stored in a forensics database for analysis.

Blockchain-based solutions were also discussed in the literature. Al-Khateeb, et al. (2019) discussed incorporating blockchain into digital incident response systems, specifically focusing on maintaining the chain of custody for digital evidence. The need for chronological documentation of evidence for the court of law was emphasised, which can be achieved using blockchain technologies.

Finally, an architecture proposed by Nanda & Hansen (2016) caters for layers introduced by cloud computing. An additional layer is introduced to the cloud computing environment that ingests forensic data. The architecture, named Forensics-as-a-service (FaaS), obtains forensic information from services running in the IaaS, PaaS and SaaS layers. The layer is operated by an external party.

The proposals discussed in this section focus on similar problem areas as the model proposed in this paper. However, no clear answer is currently documented in literature that helps to address the forensics-related information silos that currently exist between service providers. A study is therefore needed to determine a way to collect and share forensic readiness data between multiple related parties operating within service delivery environments in a secure and privacy-aware manner.

4. Cloud Computing Forensics Challenges

From a forensics perspective, the IaaS, PaaS and SaaS models are of interest as each model can be seen to introduce a layer of abstraction. IaaS provides a way to virtualise and commoditise hardware. As a result, consumers of IaaS interact with virtual representations of hardware but not the actual hardware itself. Similarly, consumers of platform offerings consume libraries hosted on IaaS. The underlying PaaS and IaaS details are rarely visible to the consumer of these platform services. Lastly, the consumer of SaaS services is generally only exposed to the applications that they consume and not to the details of the underlying platform or hosting environment. In addition, a SaaS offering may consist of PaaS and or SaaS offerings provided by multiple vendors. The inclusion of various services provided by external vendors allows companies to focus on their key strengths while outsourcing everything else to external providers.

A hierarchy, therefore, exists where IaaS is at the bottom of the hierarchy and SaaS at the top. The argument can be made that additional layering in the service model creates abstraction. At the top of the hierarchy, the layers of abstraction are so many that a user is unaware of any finer-grained hosting details. From a service provisioning perspective, the layering greatly simplifies complexities. In addition, should a breach be experienced, it becomes more challenging to perform an investigation as not all the details will be visible on all layers.

Consider the example where the entire user database of a SaaS service has been copied by an external party (adversary). The consumers of a SaaS service will typically have access to application-level audit facilities to determine if the source of the breach was a compromised user account. In some cases, an audit facility may help SaaS consumers determine if a compromised user account is to blame for a breach. However, SaaS consumers will typically not have access to any PaaS or IaaS details. For example, should an access control mechanism be broken on the PaaS layer, then it is unlikely that application-layer controls will be able to identify or compensate for such a deficiency, as the application layer has no visibility over the events occurring on the platform layer.

Similarly, should a physical breach be experienced (where a threat actor enters a data centre), then PaaS and SaaS systems are unlikely to have visibility over such events. An investigation focussing on SaaS controls, as a result, will be inconclusive. Similarly, an investigation focusing on SaaS and PaaS controls will also not produce conclusive evidence needed to identify the source of such a breach beyond a reasonable doubt. Effectively the forensic readiness capabilities of service providers form silos that are only accessible to the service providers but not to service consumers. As a result, it is argued that the layers of abstraction offered by cloud computing help for rapid service provisioning but are a burden to security and forensics.

The problems described can be formulated into a research question as follows: How can a model be defined that can help parties share information in a forensically sound manner in a cloud environment that would facilitate information sharing among service providers in a controlled manner?

The following section presents a model to help the facilitation of forensic investigations in distributed cloud environments, where multiple parties are involved that currently collect forensic data in silos.

5. Model

The lack of information sharing between parties related to an incident can potentially also have a negative impact on the incident investigation process. Each provider of a service in a cloud environment will have a view of events that transpired while providing a service to clients. An infrastructure provider will, for example, have visibility on network events, hardware failures, and physical access control breaches (among other things). Although tracked and monitored by the infrastructure provider, this information will rarely be shared with consumers of IaaS, PaaS and SaaS services. There can be many reasons for this, including privacy reasons; similarly, there will be situations in which forensic events captured by PaaS and SaaS providers are not shared with consumers of those services. Yaqoob et al. (2019) also mentioned that privacy-aware processes are generally lacking in forensic practices. The result of this is that not all information required for full transparency will ever be present unless there is a way for all the various parties to share forensically relevant information in a forensically sound manner while preserving the privacy of other tenants or users that may be involved.

A potential model for distributed forensic readiness is for each service provider in the cloud computing value chain to provide a forensic service that can be queried by upstream consumers in the SaaS value chain. Having access to forensically sound information in a controlled manner may help to reduce the information absence issue currently present in cloud environments. In addition, such a service can potentially be funded by service providers by charging service or subscription fees as an incentive.

Obtaining relevant forensic information from service providers would ensure that information not currently available for forensic investigations may, as a result, be accessible to investigators. In addition, it may also be possible to determine forensic readiness coverage of services provided internally and by external vendors, thereby helping security teams better understand their exposure to risk. Should it be possible to quantify forensic information completeness, then it may also be possible to use such a gauge to determine forensic risk (which insurers may later use to determine information security insurance premiums).

A potential problem with distributed forensic readiness would be sharing information while protecting confidential or personal information. Consider the example of two services that have been included as part of a cloud service offering, and that the service providers are in different countries. Should an incident occur and one service provider requests forensic records from the other provider, and the forensic records contain personal information, then personal information would be transferred across borders. This can be problematic if proper consent is not in place to allow this. Similar problems may also exist in terms of data governance. As a result, approval controls may be needed to ensure that distributed forensic information requests are reviewed and approved on demand as requested by external parties. Appropriate controls are needed to protect the information requested, and as a result, more steps will be needed to obtain potentially sensitive information as compared to non-sensitive information. The assumption to date has been that each service provider will operate their own forensic readiness processes in isolation from other service providers.

Responsible service providers practising forensic readiness will have controls in place to ensure that they capture all relevant information in a forensically sound manner to ensure that it will later be possible to perform forensic investigations using the information captured by their forensic readiness processes. Assuming that one cloud service consists of multiple amalgamated services provided by multiple distinct parties and forensic readiness is practised by all parties involved, then it may be beneficial for parties also involved to share forensic readiness information to ensure that a holistic picture is provided of an incident that occurred.

For example, the forensic readiness system employed by an IaaS provider will be of great use to investigators investigating a database leak of a SaaS solution if the SaaS solution provider has reason to believe that a threat actor with physical access may have been responsible for the data leak. Without sharing information, it would not be possible for a SaaS service provider to accurately determine the incident's root cause, as not all the pieces of information required will be present. By combining the pieces of forensic readiness information captured by all parties, it would be possible to obtain a clearer picture of events that occur between service providers, thereby providing a holistic picture of what happened after combining evidence from various sources.

Considering that forensic readiness information may contain personal information and considering the complexities involved with the transfer of personal information across borders, storage of personal information and processing of personal information, it can be argued that service providers are likely to retain their forensic readiness information within forensic readiness silos; in other words, service providers will capture forensic readiness information that is available to them. However, due to legal and privacy requirements, forensic readiness information will not be shared among service providers, thereby decreasing holistic visibility over forensic events.

Assuming that this model will not change anytime soon, it can be argued that there may be value in temporarily correlating forensic readiness data captured by various service providers in an attempt to understand the actions of threat actors that span across service provider domains.

A legal basis should exist for an investigator to request access to forensic services provided by other vendors. Once a vendor manages to establish a legal basis, and the legal basis is verified by the service provider with the forensic readiness artefacts of interest, then a structured discussion can occur between the two parties. An investigator will need access to very specific pieces of information, typically events that occurred within a specified timeframe. Considering that service providers are likely to service one or more clients using the same software or hardware, it can be argued that multiple layers of visibility exist, namely:

- Events that occurred for only the tenant, for example, an unauthorised user who managed to log into the supplied service application;
- Events that occurred for all tenants, for example, tenant downtime due to a failure of one of the key services relied upon by the SaaS solution;
- All events that occurred, irrespective of the tenant, for example, hardware failures on the IaaS layer, would impact all layers on top of the infrastructure layer.

A request for information for events that occurred between a specific timeframe for the tenant itself on a remote service provider will be less sensitive compared to a request for information for all events that occurred for all tenants on the same remote service provider due to the fact that sensitive information related to parties that are external to an investigation in question may be exposed to investigators. However, there may be no other option in cases where suspicion exists that multi-tenancy issues on a service provider led to a system compromise of a service hosted by a different provider.

As a result, in addition to a legal basis, consent also needs to be in place to allow investigators to provide information linked to various tenants or their systems to external investigators. In cases where consent cannot be obtained from the relevant parties involved, then some information cannot be provided to investigators. However, it will be possible to calculate a completeness score based on the consent obtained from multiple tenants involved and the amount of data represented by the tenants for which consent was obtained. Without obtaining full consent from all tenants, one can argue that there is a possibility that some pieces of a puzzle will be excluded from an investigation.

Figure 1 provides a visual depiction of components within a distributed digital forensics framework. The diagram shows a multi-tiered architecture where upper-tiered components make use of components located in lower-layered tiers. Each of the components is discussed in the following subsections.

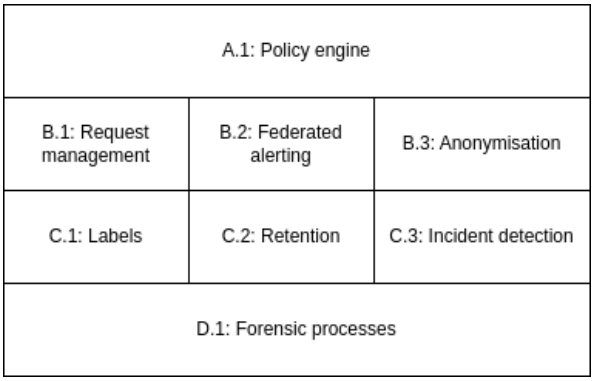


Figure 1: Federated distributed digital forensics components

5.1 Layer A.1: Policy Engine

The policy engine is the glue that holds together all underlying components in the federated model. The policy engine orchestrates all requests made in the federated environment and received in the federated environment. Predefined policies govern how external parties can access forensic artefacts, how long external parties may retain acquired evidence, controls that should be in place to protect the evidence, labels that should be applied and how requests made to external parties should be handled. An example of a policy may require a police case number, justification and special privilege levels, and manual approval before anonymised evidence can be released to a requestor. Conceptually, hundreds of policies can be defined that govern how data can be shared, anonymised and retained.

The definition of policies can potentially be done in a mark-up language, such as YAML, XML or XACML (OASIS, 2013). Since working in a federated environment, it should also be possible for a policy to reference policies applied to other federated resources.

Due to the complexities involved, defined policies will have to be validated before being applied. For federated policies referenced from remote locations, a similar validation process will have to be followed to ensure that the defined policy is valid, and that the policy is valid within the context of the referenced policy.

5.2 Layer B.1: Request Management

In a federated model, authorised parties can request forensic evidence sources relevant to the investigations performed by these parties. Requests made and received need to be validated and fulfilled in a manner that ensures nonrepudiation, confidentiality, and the integrity of data.

Request queuing occurs on this layer. It may often happen that multiple requests are submitted at once of that requests are submitted that may require manual interaction by one or more parties. As a result, requests are attached to manual and automated workflows. As part of the workflow process, requests may also include processing performed by other layers. In cases where additional permissions are requested or de-anonymised data is required, an escalation workflow can be kicked off to obtain the necessary permissions. For such a request, the number of approvers, as well as the level of authority of the approvers, will be dictated by policies managed by the policy engine.

Identity management, authentication, authorisation and platform audit logging are also handled in the request management layer.

5.3 Layer B.2: Federated Alerting

In cases where cooperation agreements are in place, it should also be possible to send alerts to federated parties when local incident detection techniques show that an event that includes one or more external service providers may be present. Alerts can be structured by considering the appropriate labels, retention and anonymisation requirements. Ultimately, alerts are structured and sent to the request management layer to process as workflow items. As a result, alerts will undergo various processing steps to ensure that alerts being delivered are in line with defined policies. This may also imply that different policies are applied for different recipients. For example, different rules can potentially be applied to recipients in Europe compared to recipients in South Africa due to differences in legal requirements.

5.4 Layer B.3: Anonymisation

Data requested by federated parties may contain sensitive information, such as keys or personal information. The sensitivity of the data in question is determined by the labels assigned. More sensitive information will require stricter controls and approvals to be in place when compared to requests containing less sensitive information. Under certain conditions, it should be possible to obtain personal information under certain conditions, but under most conditions, performing incident investigation using anonymised data would be fine up to the point where personal information may be required to identify a threat actor uniquely. By performing requests in a stepwise fashion, as described, it would be possible to perform investigations in a systematic manner while keeping the information exposed to federated parties to a minimum.

5.5 Layer C.1: Labels

Data captured in the evidence storage facility needs to be labelled to facilitate automated decision-making and anonymisation processes. Evidence labelled as containing sensitive information or personal information will require additional permissions and processing, such as anonymisation or encryption (depending on the data in

question). As part of a forensic readiness process, forensic data under the control of the service provider can be labelled according to sensitivity. Federated evidence received from external parties will contain labels, as defined by the external parties for the evidence at hand.

Sensitivity labels can be used to apply the appropriate level of anonymisation or apply the proper level of protection based on the labels being applied. In cases where local policies reference federated policies, it may, in some cases, also be required to map remote labels to local labels to ensure consistency in processing. In cases where data is received in a federated context, the labels applied by the originator should be respected.

5.6 Layer C.2: Retention

Evidence collected and received by external parties should be retained as long as is required for legal purposes. Data retention periods may often be influenced by the data labels. For example, personal information retention legal requirements may differ from the legal requirements for financial data. Similarly, federated evidence received from external parties should be kept only as long as needed to conclude an investigation and as long as permitted by relevant laws. In cases where data is received or retrieved in a federated context, the retention policies of the originator should be respected.

5.7 Layer C.3: Incident Detection

Evidence collected from various data sources needs to be correlated to detect digital incidents. The incident detection will include data collected by the service provider but can also include data obtained through the federation.

For the purposes of this discussion, an incident can be seen as any event that may be of interest to an investigator. The incident detection layer defines what an incident looks like, given the various evidence sources at hand. Minor transformations can also be defined and applied as part of the incident detection process. Defined detection definitions are to be stored in a mark-up file.

5.8 Layer D.1: Forensic Processes

Evidence captured by the service provider needs to be stored in a forensically sound manner with controls in place to ensure adequate protection against misuse. Forensic processes are concerned with the capture and preservation of forensic evidence in a forensically sound manner.

There are some technicalities related to the processing of forensic evidence obtained as part of a federated process. Software vendors in the SaaS supply chain can individually collect and preserve evidence as part of their digital forensic readiness processes. Captured artefacts can be stored within the solution provider's perimeter (referred to as local evidence storage for the purposes of this discussion). Alternatively, captured information can potentially be stored in the same storage facility used by various service providers in the SaaS value chain (such as AWS CloudTrail).

In both cases, accessed by different parties, potentially in different countries, it implies that data processing agreements need to be established among parties and that the consent of data subjects will have to be in place to perform such actions. Ultimately, such policies will be enforced by components located in higher-up tiers within the model.

The next section presents a discussion on the proposed model.

6. Discussion

The purpose of the proposed model is to help improve information sharing in environments where the sharing of information is structured in ways that respect data privacy laws. This is done through a collaborative framework that dictates the conditions under which data can be accessed.

Conceptually, the proposed model can be applied to local investigations as well as investigations that may span across organisational boundaries. The model was proposed, assuming that various forensic readiness controls that capture data on various service delivery layers will be present. The implementation and operation of controls in an organisation typically incur costs. As can be expected, organisations will have finite budgets allocated to forensic readiness, thereby implying that some forensic readiness controls will be prioritised. In contrast, others will receive little to no attention. Controls that are prioritised are typically decided through a risk assessment based on business risk. Controls that address important business risks are typically prioritised over controls that address less critical risks. In addition, there will also be risks that an organisation may not be aware of. As a result, it can be expected that controls will provide a degree of coverage, locally and remotely.

Future work will explore ways to determine the coverage of local data and remote forensic data more accurately (through federation).

It is also important to mention that organisations will adapt existing information security policies to ensure that cooperation with other organisations and law enforcement agencies is allowed.

In terms of the policy engine discussed, policies may also reference federated policies. In addition, in some cases (such as labelling), mappings between compatible labels will also have to be performed. In cases where clear misalignments exist between policies, a policy can potentially be defined that favours local policies over remote policies (or visa-versa) or requests manual intervention from an approver.

Access to forensic artefacts locally and remotely needs to be logged in a forensically accepted manner. It should be noted that federated requests should be logged in a way that provides remote parties with auditing visibility over actions taken on federated data provided to ensure accountability and traceability.

The next session will present a discussion on future work.

7. Future Work

A model to facilitate information sharing among service delivery stakeholders was discussed in this paper. Future research work will include the implementation of the model and a study of the effectiveness of the implementation of the model in real-world scenarios.

8. Conclusion

This paper describes a current issue that exists in cloud computing due to various layers of abstraction present. Service providers focus on the provision of services that include their core expertise. All other aspects of a service that is not included are typically outsourced to third parties. This can include hosting and non-core services that are included as part of a service delivery offering.

The problem identified focused on the forensic readiness silos that exist between organisations where virtually no forensic readiness information is shared among service stakeholders. The net result of this is that investigators often only have access to partial information when performing investigations. Ultimately, the lack of information can lead to inconclusive investigations due to lack of information or evidence.

A model was proposed that allows investigators to share information in a structured manner. Furthermore, the model also allows inter-organisational sharing of forensic data in a structured and privacy-aware manner to ultimately help to easy forensic investigations that may span across organisational boundaries.

There are a few discussion points that will be addressed in future research, such as the estimation of local and remote forensic readiness completeness. The implementation of a prototype will also be performed to illustrate the concept in action.

References

Alenezi, A., 2023. Digital and Cloud Forensic Challenges. *arXiv preprint arXiv:2305.03059.*

Alenezi, A., Hussein, R. K., Walters, R. J. & Wills, G. B., 2017. The Impact of Cloud Forensic Readiness on Security. *In Closer,* pp. 511-517.

Alenezi, A., Hussein, R. K., Walters, R. J. & Wills, G. B., 2017. The Impact of Cloud Forensic Readiness on Security. *In Closer,* pp. 511-517.

Al-Khateeb, H., Epiphaniou, G. & Daly, H., 2019. Blockchain for modern digital forensics: The chain-of-custody as a distributed ledger. *Blockchain and Clinical Trial: Securing Patient Data,* pp. 149-168.

Armbrust, M. et al., 2010. A view of cloud computing. *Communications of the ACM,* 53(4), pp. 50-58.

Duijn, P. A. C. & Sloot, P. M., 2015. From data to disruption. *Digital Investigation,* Volume 15, pp. 39-45.

Federici, C., 2013. AlmaNebula: a computer forensics framework for the Cloud.. *Procedia Computer Science,* Volume 19, pp. 139-146.

Kebande, V. R. & Venter, H. S., 2018. Novel digital forensic readiness technique. *Australian Journal of Forensic Sciences,* 50(5), pp. 552-591.

Lillis, D., Becker, B. & O'Sullivan, T. S. M., 2016. Current challenges and future research areas for digital forensic investigation. *arXiv preprint arXiv:1604.03850.*

Liu, F. et al., 2011. *NIST Cloud Computing Reference Architecture,* s.l.: National Institute of Standards and Technology.

Mell, P. & Grance, T., 2011. The NIST definition of cloud computing. *NIST Special Publication 800-145.*

Miller, C., Glendowne, D., Dampier, D. & Blaylock, K., 2014. Forensicloud: An Architecture for Digital Forensics Analysis in the Cloud. *Journal of Cyber Security and Mobility,* 3(3), pp. 231-262.

Monteiro, D., Yu, Y., Zisman, A. & Nuseibeh, B., 2023. Adaptive Observability for Forensic-Ready Microservice Systems. *IEEE Transactions on Services Computing.*

Nanda, S. & Hansen, R. A., 2016. *Forensics as a Service: Three-Tier Architecture for Cloud Based Forensic.* s.l., s.n., pp. 178-183.

OASIS, 2013. *eXtensible Access Control Markup Language (XACML) Version 3.0.* [Online] Available at: https://docs.oasis-open.org/xacml/3.0/xacml-3.0-core-spec-os-en.html[Accessed 22 1 2024].

Ruan, K., Carthy, J., Kechadi, T. & Crosbie, M., 2011. Cloud Forensics: An overview. *Advances in Digital Forensics VII,* pp. 36-46.

Shanmugasundaram, K., Memon, N., Savant, A. & Bronnimann, H., 2003. ForNet: A distributed forensics network. *Computer Network Security: Second International Workshop on Mathematical Methods, Models, and Architectures for Computer Network Security,* pp. 1-16.

Shetty, J., Anala, M. R. & Shobha., G., 2014. A study on cloud forensics: challenges, tools and CSP features. *Biom Bioinform,* 6(6), pp. 149-153.

Sibiya, G., Fogwill, T., Venter, H. S. & Ngobeni, S., 2013. Digital Forensic Readiness in a Cloud Environment. *Africon,* pp. 1-5.

Simou, S. et al., 2022. Revised forensic framework validation and cloud forensic readiness. *International Journal of Electronic Governance,* 14(1-2), pp. 236-263.

Villegas, D. et al., 2012. Cloud federation in a layered service model. *Journal of Computer and System Sciences,* 78(5), pp. 1330-1344.

Weir, G., Aßmuth, A. & Jäger, N., 2018. Managing forensic recovery in the cloud. *Cloud Computing.*

Yaqoob, I. et al., 2019. Internet of things forensics: Recent advances, taxonomy, requirements, and open challenges. *Future Generation Computer Systems,* pp. 265-275.

Strengthening Aviation Cybersecurity with Security Operations Centres

Wesley Murisa and Marijke Coetzee
North-West University, Potchefstroom, South Africa

wesmur@gmail.com
marijke.coetzee@nwu.ac.za

Abstract: Even though cybersecurity is a top priority for the aviation industry, research indicates that there are still many challenges to address. Modern aviation systems encompass cloud computing, OT, IoT, mobile devices, and traditional IT infrastructure. The network complexity has expanded the attack surface, leading to an increase in security incidents. Due to this complexity, detecting security incidents on time is challenging. Research indicates that it may take up to 196 days to detect an incident and another 56 days to address it, highlighting the urgency of improving security response. In this regard, establishing Security Operations Centres (SOCs) in the aviation sector must be addressed. SOCs can be instrumental in reducing the time it takes to detect and respond to security incidents. They provide visibility into threats, aid investigations, and enhance forensic efforts, enabling proactive threat mitigation. Research has been carried out on SOC implementations for specific domains like IoT, mobile devices, and higher education, neglecting aviation systems. Aviation systems such as Air Traffic Management (ATM) face unique security vulnerabilities, including signal modification, jamming, flooding, data and command injection, GPS spoofing, and blocking attacks, primarily due to their reliance on wireless technology. Most of these wireless technologies do not use encryption or authentication because they were designed to maximize performance. Insufficient funding also negatively affects ATM systems, resulting in the wide use of legacy ATM systems and a shortage of skilled personnel. ATM systems are considered critical infrastructure frequently targeted by well-resourced threat actors, including terrorists and nation-state actors, necessitating higher protection levels. This paper motivates the development of a customised SOC implementation framework for ATM systems to enhance aviation security by increasing visibility into threats and facilitating timely remediation.

Keywords: Critical infrastructure, Aviation, Air traffic management (ATM), Security operations centre (SOC)

1. Introduction

Air Traffic Management (ATM) systems have been known to be vulnerable to network security threats like signal modification and jamming, data and command injection, GPS spoofing, and blocking attacks mainly because they rely on open wireless technology (Dave *et al.*, 2022). Previously, these vulnerabilities did not pose significant safety and security challenges because ATM systems were custom-made and standalone. Today, ATM systems are massively integrated with non-aviation systems, increasing the risks of life-threatening cyber-attacks. Security vulnerabilities in ATM systems are high because civilian aviation systems do not use encryption by default to reduce operational overheads. ATM system vulnerabilities are also amplified by the failure to upgrade legacy systems in a timely manner due to the exorbitant costs of upgrading and the shortage of skilled personnel (Kagalwalla & Churi, 2019). Further, ATM systems are viewed as critical infrastructure and are often targeted by threat actors, including terrorists and well-resourced nation-state actors; hence, they require higher protection.

Security Operations Centers (SOC) have been widely recognised as an effective tool for improving cyber security in organisations, especially nowadays when the rate of cyber security threats continues to escalate across private and public sectors worldwide (Kokulu *et al.*, 2019). SOCs mainly provide visibility of security threats across the organisation, enabling forensic investigations and timely remedial action. Visibility is achieved by collecting security logs from technical controls and critical assets into a central database, which are continuously analysed to identify security incidents. Analysts are alerted when security incidents are identified. The analysts, often members of the Cyber Incident Response Team (CIRT), investigate the alerts further and take corrective action where necessary (Majid & Ariffi, 2019). SOCs can go a long way in addressing ATM cyber security challenges.

This paper is structured as follows: Section 2 discusses Air Traffic Management systems. This is followed by an outline of cyber security challenges in ATM systems in section 3. Section 4 describes SOCs before considering the current state of the art of SOCs in section 5. Section 6 discusses the recommendations for implementing SOCs for ATMs. Lastly, a concussion is given in section 7.

2. Air Traffic Management Systems

Aviation is a complex critical infrastructure system comprising different sub-systems, including ATMs, airports, and airlines (Lykou, Iakovakis & Gritzalis, 2019). ATM systems are at the core of air traffic management services. Three subsystems, namely, Communication, Navigation and Surveillance (CNS), comprise ATM systems(Lu *et al.*, 2021). Communication systems are responsible for voice and message communication between Air Traffic Control (ATC) and aircraft. In addition, pilots use communication systems to exchange information. Navigation systems are used to determine location during transit and landing. Lastly, the purpose of surveillance systems is to track an aircraft's location. The three ATM sub-systems described above, combined with Aeronautical Information Services (AIS) and meteorological services, provide information ATC uses to manage and guide aircraft (Stelkens-Kobsch, Finke & Carstengerdes, 2017). The diagram below shows how these systems are combined to give ATC a unified display system for directing and controlling aircraft.

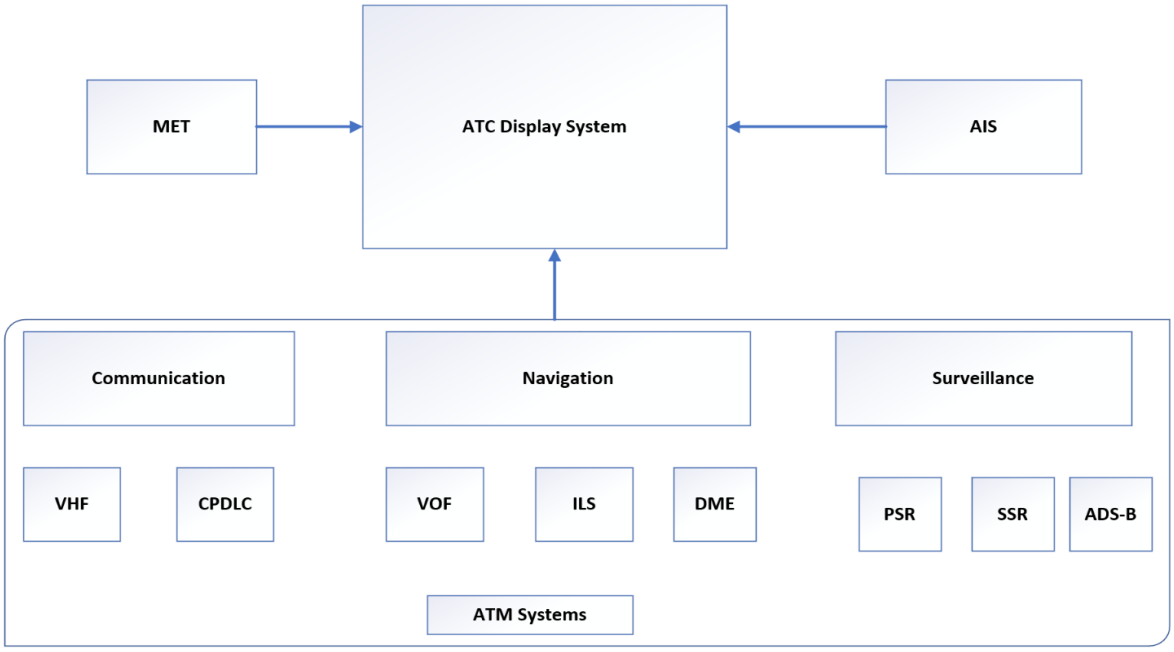


Figure 1: Air Traffic Management systems

ATM communication is based on two wireless technologies: Very High Frequency (VHF) and Controller Pilot Data Link Communication (CPDLC) systems. VHF is the primary means of communication between ATC and pilots and uses voice. It also broadcasts additional information like weather updates (Dave *et al.*, 2022). Limited range and lack of authentication are the major shortcomings of VHF. The absence of authentication on VHF makes it susceptible to denial of service and eavesdropping attacks. CPDLC is a message-based alternative to VHF that is considered more efficient, faster, and easier to use (Gurtov, Polishchuk & Wernberg, 2018). It helps reduce congestion on VHF when there are high traffic volumes and offers backup when VHF is out of range. Like VHF, it does not use authentication, making it vulnerable to several wireless attacks.

Navigation systems comprise VHF Omnidirectional range (VOR), Instrument Landing System (ILS), and Distance Measuring Equipment (DME), which are used to guide an aircraft in transit and during landing (Dave *et al.*, 2022). As the name indicates, VOR uses VHF to determine an aircraft's location relative to beacons on the ground during transit. ILS is a non-precision approach system that guides planes to the runway when conditions are unsuitable for visual landing (Sathaye *et al.*, 2019). Like VOR, ILS is fixed on the ground and uses radio signals and high-intensity light arrays to guide the pilot to the runway. DME is used with VOR to provide the pilot with navigation information during a flight.

Primary Surveillance Radar (PSR), Secondary Surveillance Radar (SSR), and Automatic Dependent Surveillance-Broadcast (ADS-B) make up surveillance systems. PSR is a passive and independent surveillance system that uses a rotating radar to detect the direction and distance of objects in airspace (Habler, Bitton & Shabtai, 2022). SSR is another surveillance system that differs from PSR in that it depends on the target aircraft to respond to its interrogative messages, for it determines an aircraft's altitude, speed, and destination (Lykou, Iakovakis & Gritzalis, 2019). PSR and SSR are similar in using radio signals to locate target aircraft details. ADS-B is the third method that is used for surveillance purposes in ATMs. In ADS-B, aircrafts continuously broadcast

their identity, location, and speed to ground stations and other aircrafts (Haass, Craiger & Kessler, 2018) . The aircraft uses the Global Navigation Satellite System (GNSS) to determine its location and velocity before sending that information to ground stations and other aircraft.

Data from the CNS systems described above is transmitted to networked computer systems, often running proprietary operating systems for storage, processing, and further transmission to the ATC display system (Lu *et al.*, 2021). Traditionally, these networks used the concept of segregation and operating system hardening to provide security against unauthorised access. Whilst segregation can be effective for providing protection, configuration errors can expose ATM networks to security threats in the local area network connected to the internet. The requirement to connect ATM systems to third parties, such as providers of meteorological data and AIS, exposes the supposedly closed network to further security risks.

The life span of ATM systems can be as long as 20 years (de Haan & Youssouf, 2023). The primary reason for the long life span is that ATM systems are critical infrastructures that are heavily regulated. New systems are subjected to rigorous validation efforts to ensure safety and reliability. In addition, the development cycles of ATM systems are lengthy and costly. Further, there is a reluctance to change critical infrastructure like ATM systems because they constantly operate, and there is little appetite for downtime. As a result, most ATM systems run on legacy systems with many vulnerabilities that malicious actors can easily exploit. Security challenges caused by legacy systems and other aspects of ATM systems are analysed in detail in the next section.

3. Air Traffic Management Security Challenges

After describing ATM systems in Section 2, this section discusses their cyber security challenges. The aviation industry is a lucrative target for cybersecurity criminal groups motivated by financial gain, as evidenced by the general increase in cybersecurity attacks over the past few years. On the one hand, terrorist groups find aviation a suitable target because successful attacks are fatal. On the other hand, national state actors also target aviation systems for intellectual property theft, intelligence gathering and political reasons. Disgruntled employees complete the list of high-profile threat actors targeting aviation systems. Therefore, ATM systems are always at risk of attack from threat actors mentioned above. The most prominent cyber security challenges in ATMs are detailed below.

Insecure by design

Most of the cyber security challenges in ATMs emanate from the lack of security considerations in the initial system design, mainly because cyber security threats were not prevalent when they were developed (Dave *et al.*, 2022). However, information security best practices require that requirements be included in the system lifecycle's design, implementation, and maintenance phase. Attempting to address security limitations after systems development is often very expensive and may introduce new security problems and system errors. The best way to address this challenge is by developing a new system that incorporates security requirements from the onset.

Unencrypted, wireless communications

Most ATM technologies use clear-text wireless communication, which cannot provide confidentiality, integrity, and non-repudiation. Although some ATM systems used by the military now use encryption, their civilian counterparts are still not encrypted due to the need to reduce operational overheads. Many ATM systems are known to be vulnerable to network security threats like eavesdropping, signal modification and jamming, data and command injection, GPS spoofing and blocking attacks mainly because they rely on unencrypted wireless technology. The availability of cheap tools such as Software Defined Radio (SDR) (Lu *et al.*, 2023) has made attacking these unencrypted wireless ATM systems easier.

Enlarged attack surface

Inherent security risks in ATM systems described above are amplified by the need to integrate aviation systems with non-aviation systems, core networks and cloud systems connected to the internet (Lu *et al.*, 2021). Air Navigation Service Providers (ANSPs), airlines, airports, and aircraft systems now share data to improve service to their tech-savvy aviation customers (Ukwandu *et al.*, 2022). The integration enlarges networks and creates new attack vectors in systems that previously relied on closed networks and proprietary protocols for security (Bernsmed *et al.*, 2022). A radical move from the view that ATM systems are closed and secure must be adopted when assessing cyber security risks.

Complex architecture

Increased interconnection and integration of aviation systems described above create a complex architecture which is difficult to secure. The result of integrations is adding external networks and their vulnerabilities to an already complex aviation sector network. Most players in the aviation sector use different technologies, including Operations Technology (OT), Internet of Things (IoT) and cloud applications supported by various supply chain partners (Kagalwalla & Churi, 2019). Other nations like the US use rail at airports, introducing an additional node to an already complex ecosystem (Szyliowicz, 2004). Although it can be argued that integrating systems improves air traffic control systems, the security risks brought about by these complexities must not be overlooked.

Legacy systems

Legacy systems, in the form of outdated technology and software, do not receive security patches once they have reached their end of life. Original Equipment Manufacturers (OEMs) and system developers often stop supporting systems that have been discontinued and as a result, use insecure protocols. This creates vulnerabilities that threat actors can easily exploit. The risks will be minimal if these legacy systems are isolated or in closed networks but as mentioned before, ATM networks are integrated with external networks. Moreover, most ANSPs are owned by national governments, which do not provide adequate funding for frequent system upgrades. As a result, most ATMS systems in use are legacy systems vulnerable to cyber security attacks.

Shortage of skilled cyber security personnel

The availability of skilled cyber security personnel in ATMs directly relates to the sector's financial situation. High demand for cyber security specialists worldwide results in the most funded sector getting the most available skills. Due to the financial challenges in the aviation sector, it isn't easy to attract and retain qualified and experienced security personnel, thereby affecting the implementation of cybersecurity programs. As a result, cyber security implementation in the aviation sector has not yet matured, although there is awareness of the importance of cyber security. This leaves the aviation sector inadequately protected and vulnerable to cyber security attacks. Statistics show that cyber security attacks in the aviation industry have increased significantly over the past few years.

Supply chain attack vectors

Other cyber risks in ATMs emanate from the supply side (Kandera *et al.*, 2022). Threat actors have devised ways of varying their attack methods by compromising aviation equipment and service suppliers. Attack vectors commonly used in aviation supply chains are software development tools, hardware components, and network connectivity. Threat actors utilise these vectors to inject malicious code into products under development or maintain remote network access. Once access has been obtained, the attackers retain it and use it to access targets in different countries serviced by that supplier. Vendors are often trusted and given remote access to their client's networks for support and maintenance. Therefore, a single attack on one supply chain partner can guarantee a threat actor access to many targets across different continents.

The challenges highlighted above expose ATMs to numerous cybersecurity threats, necessitating a comprehensive approach for mitigation. Strategies to tackle these cybersecurity challenges encompass developing secure ATM systems and the implementation of effective incident response protocols. The following section delves into the Security Operations Centre (SOC), a vital information security management tool that can enhance the overall security posture of ATMs.

4. The Security Operations Centre

SOCs are centralised security hubs established to bolster an organisation's security posture, swiftly identifying and responding to threats and security breaches before they disrupt core business operations (Majid & Ariffi, 2019). They are pivotal instruments offering comprehensive insight into security threats and attacks targeting an organization. The evolving complexity of operational landscapes, integrating diverse technologies within organizations, has made monitoring all security threats increasingly arduous (Mughal, 2022). Consequently, organizations encounter delays in recognizing and mitigating security incidents. For instance, research by Vielberth *et al.* (2020) reveals that it takes 196 days to detect an incident and an additional 56 days to address it. An effectively implemented SOC alleviates this issue by amassing logs from security controls and critical systems within the organisation, swiftly identifying threats in real-time.

SOCs are structured according to the People, Processes, and Technology (PPT) framework (Vielberth *et al.*, 2020) with Governance constituting another critical element that can be incorporated into the PPT framework.

People encompass personnel from various domains, including business, information technology (IT), and security analysts engaged in establishing and managing the SOC. Technology pertains to the technical tools deployed for data collection, analysis, and issuing security alerts upon detecting attacks. Governance and processes, on the other hand, encompass incident response playbooks, policies, procedures, and incident management protocols.

Security logs from technical controls and critical assets are funnelled into a Security Information and Event Management (SIEM) system to achieve visibility. The logs from all onboarded systems undergo ingestion, normalization, correlation, and analysis to pinpoint suspicious security events. Analysts receive notifications when such events are detected via the service desk system, and subsequent investigations determine whether the suspected events are true or false positives (Saraiva & Mateus-Coelho, 2022). False positives are closed within the service desk system, while true positives are subject to further investigation and remediation. Policies and procedures define the types of security incidents that analysts can handle independently and those that require a CIRT.

Maintaining visibility over threats and attacks targeting organisations is paramount, especially considering the vast array of technologies employed, ranging from cloud and on-premises systems to mobile and IoT devices (Miloslavskaya, 2016). The diversity of technologies, including ATMs, introduces complexities in monitoring and managing them effectively without the aid of a SOC. A SOC bolsters an organisation's security stance by enabling the centralized tracking of diverse technologies, even if they are geographically dispersed.

While the primary function of a SOC is to provide visibility into security threats, it also performs other essential functions. SOCs are crucial components of Cyber Security Incident Response Teams (CSRITs), aiding in threat detection, aggregating threat information across the network, and supporting forensic investigations. Another function of a SOC is to monitor security controls deployed within an organisation(Jacobs, Arnab & Irwin, 2013). This monitoring function enables organisations with a SOC to gain a centralized view of all security threats detected by security controls such as firewalls, Intrusion Prevention Systems (IPS), and antimalware systems. The SOC promptly detects the unavailability of any onboarded technical controls. Organizations are often legally mandated to retain logs for specific periods, a function efficiently fulfilled by a SOC. The UK government's Her Majesty's Government (HMG) mandates the establishment of a SOC (Onwubiko & Ouazzane, 2019). SOCs can also assist in meeting various compliance requirements, including laws and regulatory frameworks like GDPR, the POPI Act, PCI DSS, and ISO 27001.

SOCs are categorized as internal when operated by an organization's employees and external when managed by external service providers (Onwubiko, 2015). Many organizations opt for a hybrid model, where external service providers handle monitoring and alerting while the host organisation's IT engineers manage remediation functions. Challenges in attracting and retaining qualified cybersecurity experts may influence the choice between internal and external SOC models in attracting. Internal and external SOCs can be hosted on-premises or in the cloud.

Concerning architecture, organizations venturing into SOC establishment can choose between centralized, decentralized, and distributed architectures (Vielberth *et al.*, 2020). In a centralized SOC, logs from different locations or subsidiaries are sent to a central hub for processing. In the distributed architecture, SOCs are deployed in various locations or subsidiaries but appear as one integrated SOC to end users. Decentralized SOCs encompass a combination of distributed and centralized elements. They deploy SOCs with limited functionalities in diverse locations or subsidiaries, forwarding their data to a central SOC for processing. Specific business requirements should guide the choice of SOC architecture.

While much research has been conducted on SOCs, a limited focus has been on ATM systems. Existing research has explored SOC implementations for IoT and mobile devices (Suomalainen et al., 2022), SOC implementations in higher education (Gamilla & Palaoag, 2022), and visualization techniques to enhance SOC performance (Mihindu & Khosrow-shahi, 2020). Other research has aimed to improve SOC effectiveness by addressing critical asset onboarding and identifying challenges (Onwubiko, 2021). Notable systematic literature reviews (Schlette, Vielberth & Pernul, 2021) have generated recommendations for enhancing SOC effectiveness. Additionally, research has extensively explored ways to improve SOC analyst performance (Agyepong *et al.*, 2020). Jacobs, Arnab and Irwin (2013) has proposed a framework for measuring SOC services' effectiveness while Villalón-Huerta, Gisbert and Ripoll-Ripoll (2022) introducing a technology-independent defensive kill chain for incident response to detect and respond to threats.

5. Current State of the art on Aviation SOCs

After examining the role of Security Operations Centers (SOCs) in enhancing visibility into cybersecurity threats in ATMs, this section explores exemplary ATM SOCs worldwide that can serve as models for the African continent. SOCs support cybersecurity incident response and information-sharing initiatives in Europe and the United States. However, the implementation of SOCs in ATMs is not formally documented in academic research. Detailed insights into SOC utilization are provided below.

EUROCONTROL established the European Air Traffic Management Computer Emergency Response Team (EATM-CERT) to proactively detect, prevent, respond to, and recover from cybersecurity attacks targeting ATM systems in Europe (Mana & Friligkos, 2019). EATM-CERT operates SOCs at local, national, and regional levels to bolster incident response teams (Lekota & Coetzee, 2021). EATM-CERT collaborates with national European CERTs to share relevant cybersecurity information related to ATMs. Additionally, EATM-CERT lends its expertise to various national CERTs to manage ATM-related cybersecurity incidents effectively. EATM-CERT employs a Malware Information Sharing platform to disseminate ATM cybersecurity threat intelligence across multiple European countries. Furthermore, EATM-CERT helps its members by offering guidelines for procuring SOC services and addressing the cybersecurity skills gap in ATM services.

In response to escalating cybersecurity threats against aviation systems in the US, the Aviation Information Sharing Center (A-ISAC) was established (A-ISAC, 2023). A-ISAC's membership includes airlines, airports, Original Equipment Manufacturers (OEMs), government CERTs, and ANSPs, all actively engaged in threat detection, prevention, and remediation. This collaboration involves sharing vulnerabilities, threat intelligence, policies, and standards on a secure, common platform. Each stakeholder operates its own SOC, gathering threat information shared among peers. Additionally, A-ISAC monitors the dark web and social media for additional threat intelligence. A-ISAC fosters information exchange through forums, including working groups and summits, where aviation cybersecurity experts convene to discuss critical topics.

The EATM-CERT and A-ISAC exemplify cutting-edge SOC implementations in addressing cybersecurity challenges in ATMs. EATM-CERT focuses on ATM-specific solutions, while A-ISAC caters to the broader aviation industry. Nevertheless, both approaches provide valuable insights for establishing effective SOCs in African ATM systems. The subsequent section outlines recommendations for implementing SOCs in African ATMs.

6. Security Operations Centre for Air Traffic Management

Addressing cybersecurity vulnerabilities in critical infrastructures such as ATMs is crucial to ensuring safety and safeguarding human lives. The urgent attention needed to tackle the lack of security in most ATM systems cannot be overstated, as successful attacks on these systems can have fatal consequences. While efforts are underway to develop more secure ATM systems, the lengthy development cycles and rigorous testing protocols necessary before implementation mean that existing security flaws will persist for some time. This underscores the pressing need to confront cybersecurity challenges in ATMs. SOC capabilities can address ATM challenges directly and indirectly as summarised in the table 1 below.

Sophisticated and resourced threat actors already target ATM systems because they are critical infrastructure. Security challenges such as legacy systems, insecure by design, complex architecture, enlarged attack surface and supply chain attack vectors, make ATM systems more vulnerable to cyber security threats. These challenges are directly addressed by a SOC through automated threat monitoring. Automated threat monitoring provides visibility of the attacks and supports incident response teams. Auto remediation rules set in SOCs resolve security incidents as soon as they are identified without manual intervention. Furthermore, storing security logs aligns with the compliance requirements of different security governance frameworks such as NIST and support forensic investigations.

In SOCs, all critical systems and security control send logs to the SIEM to detect security threats, as described in Section 4. Onboarding these systems provides a clear picture of the systems in use and structure of ATM networks, thereby indirectly addressing the complex architecture challenge described in section 3. Documenting systems in use and their relationships can be easily achieved when planning the systems to be onboarded to a SOC because all systems are identified and documented. Network documentation can be helpful for evaluating and simplifying the network structure. In addition, security alerts from the SOC will provide insight into which part of the network is most vulnerable and requires urgent attention.

Implementing SOCs in ATMs also forces the sector to indirectly address skilled cyber security manpower challenges. SOCs require skilled cyber security personnel. Therefore, management is forced to find a solution to this problem. The problem can be addressed in different ways, including outsourcing, training, headhunting and attracting skilled personnel through lucrative remuneration. Despite the method chosen to address the

skill shortage, the sector benefits from the availability of skilled and experienced personnel. SOC functions such as automated remediation of security threats can be handy in addressing human resources challenges. This will result in the maturity of cybersecurity management programs and reduce security threats in the long run.

While implementing a SOC in ATM systems can go a long way in addressing cybersecurity challenges in the sector, the costs involved are huge and underfunded ANSPs may struggle to finance setting up a full-fledged SOC all at once (Vaarandi & Mases, 2022). SOC building blocks, such as building the cyber security skill base and implementing an SIEM, can be achieved relatively cheaply before an entire SOC can be set up. While these two steps will not provide as much benefit as setting up a SOC, they are critical components of a SOC that provide some protection while preparing for full SOC implementation.

An additional impediment to effectively addressing ATM cyber security challenges through SOCs lies in the limited knowledge and awareness of cyber security challenges by aviation stakeholders. A survey by (Strohmeier *et al.*, 2019) reveal a gap in appreciation of the potential impact of cyber security on Air Traffic Control (ATC) technology although some progress has been noted in Europe. The authors characterise aviation stakeholders response to cybersecurity issues as defensive and secrecy. Consequently, it is imperative to bridge the cybersecurity knowledge gap among stakeholders to ensure the viability of SOCs as a solution to ATM cyber security challenges.

Table1: Mapping of ATM cyber security problems to solutions provided by SOCs

ATM security challenges	SOC Solutions
Insecure by design	Automated threat monitoring Auto-remediation Support for incident response activities and forensic
Unencrypted, wireless communications	
Enlarged attack surface	
Legacy systems	
Supply chain attack vectors	
Shortage of skilled cyber security personnel	Auto-remediation and Outsourcing
Complex architecture	Network Documentation

Although research on SOCs has been conducted in various contexts, such as mobile devices and higher education, ATM systems pose unique challenges as critical infrastructures that blend Operations Technology (OT) systems like PSR with computer networks to provide services to ATC. The ATM network frequently integrates with external systems such as meteorological information service providers and AIS. Additionally, systems like CPDL expand ATM networks to aircraft and airline networks. The attack surface is further broadened by providing remote support and maintenance access to supply chain partners. Further, cloud systems and mobile devices are integrated into the same networks. This calls for a unique approach to obtaining the best out of using SOCs to curb cybersecurity threats in ATMs.

7. Conclusion

ATM systems are vulnerable to cyber security threats from different actors motivated by different reasons. Security challenges described in section 3 make it easier for threat actors to exploit ATM systems vulnerabilities, but SOCs can provide threat visibility and support incident response activities. While lessons can be learned from the current state of the art described in section 5, the African continent does not have fully functional cyber security structures on a continental level, as in the US and Europe. A formalised approach that guides the implementation of SOCs for ATM systems is required, as has been done for other sectors like higher education. Such an approach should consider the ATM system-specific challenges like legacy systems, complex architectures, shortage of skilled manpower, inadequate funding and enlarged attack surface. Africa can learn from the US and Europe by pulling resources and establishing SOCs at a continental level. Skilled cyber security experts can be employed at a continental level and support all member states. Such an initiative can go a long way in reducing the prevalence of security incidents on ATM systems on the continent.

References

A-ISAC. 2023. *A-ISAC* https://www.a-isac.com/ Date of access: 05 November 2023.

Agyepong, E., Cherdantseva, Y., Reinecke, P. & Burnap, P. 2020. Towards a framework for measuring the performance of a security operations center analyst. In. 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security), Dublin, Ireland. pp. 1-8.

Bernsmed, K., Bour, G., Lundgren, M. & Bergström, E. 2022. An evaluation of practitioners' perceptions of a security risk assessment methodology in air traffic management projects. *Journal of Air Transport Management*, 102:102223. 10.1016/j.jairtraman.2022.102223

Dave, G., Choudhary, G., Sihag, V., You, I. & Choo, K.-K.R. 2022. Cyber security challenges in aviation communication, navigation, and surveillance. *Computers & Security*, 112:102516. 10.1016/j.cose.2021.102516

de Haan, J. & Youssouf, A. 2023. Cryptography Based Security for the ATM Surveillance Chain. In. 2023 IEEE/ACM 4th International Workshop on Engineering and Cybersecurity of Critical Systems (EnCyCriS), Melbourne, Australia. IEEE. pp. 31-36.

Gamilla, A.P. & Palaoag, T.D. 2022. Building a Barrier: A security operations center framework for a sustainable smart campus network. In. 2022 6th International Conference on Information Technology (InCIT), Nonthaburi, Thailand. pp. 256-261.

Gurtov, A., Polishchuk, T. & Wernberg, M. 2018. Controller–Pilot Data Link Communication Security. *Sensors*, 18(5):1636. 10.3390/s18051636

Haass, J.C., Craiger, J.P. & Kessler, G.C. 2018. A Framework for Aviation Cybersecurity. In. NAECON 2018 - IEEE National Aerospace and Electronics Conference, Dayton, OH, USA. pp. 132-136.

Habler, E., Bitton, R. & Shabtai, A. 2022. Evaluating the Security of Aircraft Systems. *arXiv preprint arXiv:2209.04028*, 10.48550/arXiv.2209.04028

Jacobs, P., Arnab, A. & Irwin, B. 2013. Classification of security operation centers. In: HS Venter, M.L.a.M.C., ed. 2013 Information Security for South Africa, Sandton, South Africa. pp. 1-7.

Kagalwalla, N. & Churi, P.P. 2019. Cybersecurity in aviation: An intrinsic review. In. 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), Pune, India. pp. 1-6.

Kandera, B., Holoda, Š., Jančík, M. & Melníková, L. 2022. Supply chain risks assessment of selected EUROCONTROL's surveillance products. In. 2022 New Trends in Aviation Development (NTAD), Novy Smokovec, Slovakia. pp. 86-89.

Kokulu, F.B., Shoshitaishvili, Y., Soneji, A., Zhao, Z., Ahn, G.J., Bao, T. & Doupé, A. 2019. Matched and mismatched SOCs: A qualitative study on security operations center issues. In. Proceedings of the ACM Conference on Computer and Communications Security, London United, Kingdom. pp. 1955-1970.

Lekota, F. & Coetzee, M. 2021. Aviation Sector Computer Security Incident Response Teams: Guidelines and Best Practice. In. European Conference on Cyber Warfare and Security, Online. Academic Conferences International Limited. pp. 507-XII.

Lu, X., Dong, R., Wang, Q. & Zhang, L. 2023. Information Security Architecture Design for Cyber-Physical Integration System of Air Traffic Management. *Electronics (Switzerland)*, 12(7), 1665. 10.3390/electronics12071665

Lu, X., Wu, Z., Wu, Y., Wang, Q. & Yin, Y. 2021. Atmchain: Blockchain-based solution to security problems in air traffic management. In. 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), San Antonio, TX, USA. pp. 1-8.

Lykou, G., Iakovakis, G. & Gritzalis, D. 2019. Aviation cybersecurity and cyber-resilience: assessing risk in air traffic management. *Critical Infrastructure Security and Resilience: Theories, Methods, Tools and Technologies*:245-260. 10.1007/978-3-030-00024-0_13

Majid, M. & Ariffi, K. 2019. Success factors for cyber security operation center (SOC) establishment. In. Proceedings of the 1st International Conference on Informatics, Engineering, Science and Technology, INCITEST 2019, 18 July 2019, Bandung, Indonesia, Bandung, Indonesia.

Mana, P. & Friligkos, V. 2019. Eurocontrol/Eatm-Cert Services - Supporting Aviation To Better Manage Cyber Threats. In. 2019 Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, VA, USA. pp. 1-15.

Mihindu, S. & Khosrow-shahi, F. 2020. Collaborative visualisation embedded cost-efficient, virtualised cyber Security operations centre. In. 2020 24th International Conference Information Visualisation (IV), Melbourne, Australia. pp. 153-159.

Miloslavskaya, N. 2016. Security operations centers for information security incident management. In. 2016 IEEE 4th International Conference on Future Internet of Things and Cloud (FiCloud), Vienna, Austria. pp. 131-136.

Mughal, A.A. 2022. Building and Securing the Modern Security Operations Center (SOC). *International Journal of Business Intelligence and Big Data Analytics*, 5(1):1-15.

Onwubiko, C. 2015. Cyber security operations centre: security monitoring for protecting business and supporting cyber defense strategy. In. 2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (cybersa), London, UK. pp. 1-10.

Onwubiko, C. 2021. Rethinking security operations centre onboarding. In. 2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), Dublin, Ireland. pp. 1-9.

Onwubiko, C. & Ouazzane, K. 2019. Challenges towards building an effective cyber security operations centre. *International Journal on Computational Science & Applications*, 4(1):11-39. 10.48550/arXiv.2202.03691

Saraiva, M. & Mateus-Coelho, N. 2022. CyberSoc Framework a Systematic Review of the State-of-Art. *Procedia Computer Science*, 204:961-972. 10.1016/j.procs.2022.08.117

Sathaye, H., Schepers, D., Ranganathan, A. & Noubir, G. 2019. Wireless attacks on aircraft landing systems. In. Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks, Miami Florida. pp. 295-297.

Schlette, D., Vielberth, M. & Pernul, G. 2021. CTI-SOC2M2–The quest for mature, intelligence-driven security operations and incident response capabilities. *Computers & Security*, 111:102482. 10.1016/j.cose.2021.102482

Stelkens-Kobsch, T.H., Finke, M. & Carstengerdes, N. 2017. A comprehensive approach for validation of air traffic management security prototypes: A case study. In. 2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC), St. Petersburg, FL, USA. pp. 1-10.

Strohmeier, M., Niedbala, A.K., Schäfer, M., Lenders, V. & Martinovic, I. 2019. Surveying aviation professionals on the security of the air traffic control system. In. Security and Safety Interplay of Intelligent Software Systems: ESORICS 2018 International Workshops, ISSA 2018 and CSITS 2018, Barcelona, Spain, September 6–7, 2018, Revised Selected Papers, Barcelona, Spain. pp. 135-152.

Szyliowicz, J.S. 2004. Aviation security: promise or reality? *Studies in conflict & terrorism*, 27(1):47-63. 10.1080/10576100490262160

Ukwandu, E., Ben-Farah, M.A., Hindy, H., Bures, M., Atkinson, R., Tachtatzis, C., ... Bellekens, X. 2022. Cyber-security challenges in aviation industry: A review of current and future trends. *Information*, 13(3):146. 10.3390/info13030146

Vaarandi, R. & Mases, S. 2022. How to build a SOC on a budget. In. Proceedings of the 2022 IEEE International Conference on Cyber Security and Resilience, CSR 2022, Rhodes, Greece. pp. 171-177.

Vielberth, M., Bohm, F., Fichtinger, I. & Pernul, G. 2020. Security Operations Center: A Systematic Study and Open Challenges. *IEEE Access*, 8, 10.1109/ACCESS.2020.3045514

Villalón-Huerta, A., Gisbert, H.M. & Ripoll-Ripoll, I. 2022. SOC Critical Path: A Defensive Kill Chain Model. *IEEE Access Volume 10*, 10:13570-13581. 10.1109/ACCESS.2022.3145029

Quantum-Secure Signalling Model for L1/L2 Next-Gen Interconnect and Roaming Networks Over IPX for NB-IoT Traffic: A Review

Themba Ngobeni and Boniface Kabaso
Cape Peninsula University of Technology, Cape Town, South Africa

216187133@mycput.ac.za
kabasob@cput.ac.za

Abstract: Signalling security is critical for next-generation mobile (NGN) networks to ensure integrity, privacy, and confidentiality of communication protocols. However, recent research on existing 5G standards has uncovered vulnerabilities that could be exploited by emerging surveillance threats such as HiddenArt and Harvest Now Decrypt Later (HNDTL), as well as the prospect of quantum computing threatening to break classical cryptographic techniques. This systematic literature review (SLR) investigates quantum-based signalling solutions to mitigate these threats in interconnect and roaming 5G networks (IRN) over IP eXchange for Narrowband Internet of Things (NB-IoT) traffic. A database search was conducted, studies were selected and compared based on methods, approaches, techniques, and limitations. The SLR found that quantum key distribution (QKD) combined with quantum teleportation (QT) can protect and mitigate threats and attacks on signalling networks when protocols interact and interoperate in carrier networks. QKD techniques can be used for L1 protocols, while secure direct quantum communication can be used for L2 protocols. This study concludes that further studies are needed to integrate different techniques and protocols for L1/L2 signalling networks to create a robust quantum-secure signalling model for global interconnect and roaming NGN.

Keywords: Quantum teleportation, Quantum cryptography, Interconnect and roaming, NB-IoT, L1/L2 Signalling, Next-generation networks

1. Introduction

Signalling security is critical for the NGN to ensure integrity, privacy, and confidentiality of communication protocols (ITU-T, 2021). Signalling, the mechanism used to control network communications and network management, serves as a crucial attack surface for misuse and unauthorized access in wireless and mobile telecommunications (ITU-T, 2021). It also serves as the primary source of network intelligence, analytics, and user behaviour monitoring (Baraković Husić et al., 2012). Exploiting mobile signalling infrastructure can turn the network into a tool for cyber warfare when used by malicious actors (COTS Staff, 2023). Additionally, emerging threats like HNDTL often goes undetected in current L1 signalling networks (Garcia Cid et al., 2022; Cho, 2019). ENEA (COTS Staff, 2023) describes signalling security in mobile communications as a systematic blind spot, emphasising the need for improvements in IRN Signalling security. Roaming enables Mobile Network Operator (MNO) to access services remotely through Visited Public Mobile Networks (VPMN) from their Home Public Mobile Network (HPMN) (GSM Association, 2021). To use VPMN services, a roaming agreement between HPMN and VPMN is essential, with standardized commercial procedures (GSMA, 2020; Lutu, Perino, et al., 2021). Signalling security in the context of IPX pertains to vulnerabilities in telecommunication protocols used for roaming interworking (ENISA, 2018). With the increasing deployment of commercial and industrial Internet of Things (IoT) devices, the need for permanent roaming connectivity is growing (Lutu, Trevisan, et al., 2021).

The advent of 5G represents the NGN, featuring enhanced usecases such as higher bandwidths, lower latencies, and higher communications (C. Yu et al., 2021; ITU-R, 2015; Manzalini, 2020). The adoption of IoT over 5G, specifically 5G NB-IoT, offers cost-effective, low-rate data connectivity (Tan et al., 2022). Additionally, outlines that a significant portion of IoT applications are expected to run on NB-IoT deployment mode by the end of 2025. However, enhanced usecases introduce security challenges that traditional cryptographic solutions cannot effectively address due to stringent latency requirements (Djordjevic, 2022; Dutta, 2022). The abuse of telecommunications infrastructure by surveillance spyware further exacerbates the challenges faced by IRN signalling traffic (Holtrup et al., 2021). The United Nations recognizes the importance of security in connectivity and communication as a human right (United Nations, 2016). To address these critical issues, the International Telecommunication Union (ITU) organized a workshop aimed at improving signalling protocol security (ITU-T, 2021). MNOs have also underscored the importance of secure interconnection protocols, such as Security Edge Protection Proxy (SEPP) for 5G roaming (Intelligence & IBASIS, 2020). The challenges facing NGN's IRN underscore the necessity of secure communication protocols to address threats such as quantum computing, HNDTL, eavesdropping, interception, and key exchange difficulties in control plane signalling and user data traffic. Quantum communication (QC) offers a solution to these challenges as it is immune to tampering, eavesdropping, and interception communication (Pirandola et al., 2015; Asfaw et al., 2022). Implementing unconventional protocols in the IRN can improve signalling security for the NGN over IPX, providing enhanced

security for roaming subscribers. In this study, we explore the pressing need for improved signalling security in NGN, the challenges faced, and promising solutions, including the use of QC to enhance the security of roaming subscribers in NGN over IPX for NB-IoT traffic.

2. Related Work

The literature reveals critical concerns within the realm of mobile network security and highlights the inadequacies in addressing these issues. Notably, it is emphasized that MNOs have not given sufficient attention to security within the IRN signalling interfaces (Køien, 2021; Tezergil & Onur, 2021). In addition, the transition to IP-based transport networks is deemed unreliable and lacking Quality of Service (QoS) guarantees (Baraković Husić et al., 2012). The absence of end-to-end security, integrity protection, and encryption in signalling further exacerbates security vulnerabilities (Tan et al., 2022; You et al., 2021). As 5G technology evolves, it faces challenges related to ensuring high computing and communication requirements, ultimately impacting QoS (You et al., 2021; Køien, 2021). Consequently, there is a pressing need to address these security issues to safeguard both subscribers and the networks (Abdel Hakeem et al., 2022). Furthermore, the literature underscores the necessity for continuous development and improvement in security measures for mobile networks, as advocated by 3GPP, including enhancing the security of both control and user plane signalling messages (Abdel Hakeem et al., 2022; Garzon et al., 2022; Køien, 2021).

The emerging 5G technology poses unique vulnerabilities that require novel approaches and technologies for mitigating security risks (Kim, 2020; Chih-Lin et al., 2016). In light of these concerns, the integration of quantum-based technologies, particularly Quantum Communication (QC) and Quantum Cryptography (QCrypt), is presented as a promising avenue for addressing these security challenges (Wang et al., 2020). QC offers unconditional security, as eavesdropping attempts are detectable, making it a valuable asset in safeguarding data transmission (Ngobeni et al., 2020; Djordjevic, 2022). However, as quantum-assisted crypto analysers pose a new threat, the security community must focus on devising post-quantum techniques and standardized security protocols (Suomalainen et al., 2018). While the literature covers various aspects of quantum-based technologies, it underscores the lack of quantum-based signalling models for interconnect and roaming networks over IPX, emphasizing the significance of developing a comprehensive quantum-based signalling model for NB-IoT traffic over IRN. In summary, the literature highlights the pressing security challenges faced by NGN, particularly in the context of 5G technology. The inadequacies in addressing these issues have spurred the exploration of quantum-based solutions to enhance the security of signalling interfaces and roaming networks over IPX, aiming to create robust signalling L1/L2 models against evolving threats.

3. Research Aim and Objectives

This research aims to find existing quantum-based security solutions for protecting L1/L2 signalling traffic in the IRN over IPX and compare them in terms of approach, method, technique, and constraints. The main objectives are:

- To review existing body of knowledge on quantum-based security solutions to protect interconnect and roaming signals.
- To review how the identified quantum-based signalling solutions compare with each other in terms of technology, technique, approach, methods, and constraints.
- To identify strength of evidence supporting the various quantum-based signalling techniques and how they can be best applied to develop a secure signalling model, and
- To outline the implications of the research to explore new ideas and possibilities.

4. Methodology

SLR was adopted with features of the PRISMA statement (Moher et al., 2009) and followed the process outlined by (Okoli, 2015). SLR is a beacon of clarity that provides a rigorous and transparent methodology for synthesising existing research. This study argues that SLR stands out among other research methodologies for its unique ability to provide unbiased, repeatable, comprehensive, and cumulative evidence.

4.1 Review Protocol

The research questions (RQs) posed for this review are: *RQ1*: What existing quantum-based security solutions have been developed for NGN-enabled NB-IoT traffic?, *RQ2*: How do the various quantum-based signalling solutions found in answer to RQ1 differ in terms of technology, technique, approach, method, and limitations?, *RQ3*: How strong is the evidence in support of the different quantum-based signalling security solutions? and

RQ4: What are the potential implications of these findings of the research when developing a new quantum-based signalling security solution for NGN-enabled IoT over IPX?

4.2 Search Query

For comprehensive and relevant SLR, we were in favour of curated databases such as IEEE Xplore, arXiv, ScienceDirect, ProQuest and ACM Digital Library. They offer precise searching, trustworthy research and seamless analysis tools to drive research forward. Search terms applied: quantum* AND (communication OR teleportation OR entanglement OR QKD) AND (roaming OR interconnect OR security OR "physical layer" OR "data link Layer" OR "L1/L2" OR L1 OR L2) AND (mobile OR wireless OR IoT OR IPX). All searches conducted from the period of 2018 to 2023.

4.3 Study Selection

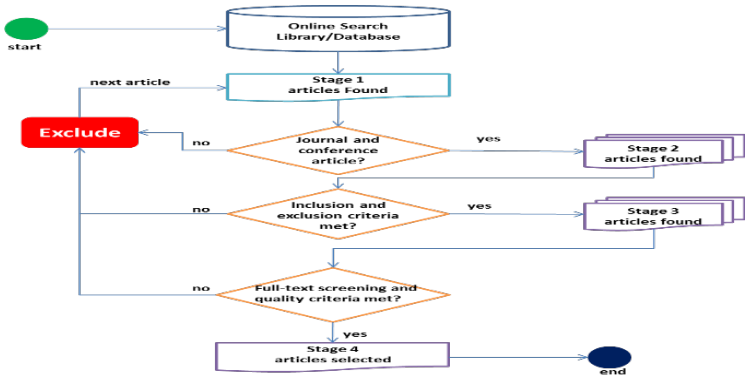


Figure 1: Execution process

Figure 1 outlines the primary study protocol, which includes four main stages: the initial search, the identification of relevant literature, the application of inclusion and exclusion criteria, and a quality assessment. The review aims to identify the current state of knowledge in the field of quantum-based signalling security, which includes disciplines such as quantum communication, quantum mechanics, computer science, and information and communication technology (ICT).

Inclusion criteria:

- Peer-reviewed empirical studies obtained online and open access.
- Empirical studies with advances in quantum security for NGN's IRN.
- Studies that evaluate the security, effectiveness, or scalability of quantum communication.

Exclusion criteria:

- Articles not related to quantum communication and not addressing L1/L2 WAN protocols.
- Studies not written in English, and studies before 2018.

4.4 Quality Assessment (Qa) Criteria

We followed a peer-reviewed journal articles and conference proceedings. Each paper has been carefully assessed to ensure quality. The study's abstract, methodology and conclusion were reviewed for applicability and quality. The quality score indicated by 1 for agree, 0.5 partial and 0 disagree, any score less than 6 out of 10 is rejected. Table 1, shows the questions used for the QA.

Table 1: Quality Assessment Questions

Qua#	Quality Assurance Questions
Q0	Is the aim of this research study clearly stated?
Q1	Is the quantum security method and approach clearly defined?
Q2	Is the transmission technique and algorithm used clearly defined?
Q3	Does the protocol support L1/L2 WAN infrastructure?
Q4	Does the study address interconnect and roaming or IPX?
Q5	Is the data collection procedure relevant for L1/L2 security?

Qua#	Quality Assurance Questions
Q6	Does the results and findings contribute to signalling security?
Q7	Are the limitations specified?
Q8	Is it a repeatable research methodology?
Q9	Does study contribute to the telecommunications?

4.5 Data Extraction Form

Table 1: Data Extraction Form

Variables	Purpose
SID	to uniquely identify the study for the article
Authors	to represent the article authors
Title	to capture the title/topic of the research article/journal
Year	to capture publication year
OSI layer	to indicate which OSI layer used L1/L2
Quantum Approach	Quantum communication Approach applied
Quantum Technology	Physical implementation used
Quantum Technique	Quantum method/technique applied
Method/Protocol	to indicate the protocol used
WAN infrastructure	WAN infrastructure supported (Optical, Freespace, Satellite, etc.)
Quality of Evaluation	Evaluation score (out of 10)
Strengths	to capture article strength
Weaknesses	to capture articles weakness
Gap	To identify a gap or gaps in the study

4.6 Data Synthesis

The data collected was not statistically analysed. However, the selected study in this report presents key facts using a narrative and tabular format. Combining narrative and tabular formats is an effective strategy to achieve a comprehensive and impactful presentation. Our choice is guided by the research objectives, QA criteria and target audience. Since clarity, accessibility and engagement are paramount, narrative and tabular formats offer valuable advantages over purely statistical formats.

5. Findings and Discussion

RQ1: What existing quantum-based signalling security solutions have been developed for NGN-enabled NB-IoT traffic?

Table 2: Process Results [2018-2023]

Stage	Description	Studies#	Excluded
0th	Online library/database search	2708	-
1st	Filter by Journal or article	134	2574
2nd	Filter by Inclusion, exclusion criteria	92	42
3rd	Full text gathered	59	33
4th	Selected study	13	46

To answer the research questions, the studies were briefly summarised using a yearly breakdown, see Fig. 2.

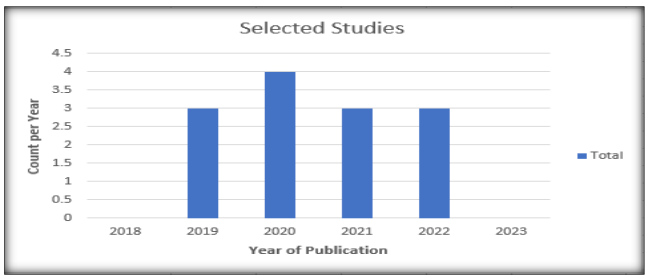


Figure 2: Publication on quantum-based security solutions

In this review, 13 key studies were selected. Several quantum-based communication solutions in the literature were examined, which encompasses aspects of the development and application of quantum communication technologies. The selected studies have been given an identifier (SID) followed by citation e.g S#[<ref>] as follows: S1(Kozlowski et al., 2020), S2(Abulkasim et al., 2019), S3(Chen et al., 2020), S4(Khawasik et al., 2022), S5(Li et al., 2019), S6(Li et al., 2022), S7(Lou et al., 2019), S8(Ntanos et al., 2021), S9(Pan et al., 2020), S10(N. Yu et al., 2021), S11(Zimmer et al., 2021), S12(Tanizawa & Futami, 2020), S13(Tanizawa & Futami, 2022). Four categories were identified and divided into quantum cryptography, quantum teleportation, quantum networks and other areas of quantum communication. All these studies contribute to the advancement of quantum communication and provides new insights into the challenges and opportunities. The studies have various level of application in terms of open system interconnection (OSI) physical layer (L1) and datalink layer (L2) categorised in terms of techniques and protocols on how they lead to signalling security enhancements for L1 and L2 as:

OSI L1:

- Quantum cryptography (Qcrypt) uses quantum mechanics principles to secure communication, considered more secure as it's impossible to eavesdrop without being detected. It can authenticate devices, encrypt signalling traffic and create secure key exchange: *S4, S8, S9, S12, S13.*
- Quantum teleportation (QT) uses quantum entanglement to transfer the quantum state to other locations. This can be used to protect signalling messages by ensuring that it cannot be intercepted or modified in transit: *S3.*
- Quantum networks (QN) uses quantum technologies to transmit and process information. This can be used to protect signalling messages by providing a secure and reliable communication channel: *S1, S3, S6, S10, S11.* Other quantum communication protocols *S2, S5 and S11.*

OSI L2:

- Quantum secure direct communication (QSDC) allows two parties to communicate directly without need for a shared key using QT. This can be used to protect signalling on OSI L2 by ensuring signalling messages cannot be intercepted or eavesdropped: *S5* and *S9*.
- QN's security of signalling using packet level quantum network intercommunication: *S10.*
- Quantum signature allows two parties to sign messages in a secure manner, this can be used to protect the authenticity and integrity of signalling messages using Qcrypt: *S7* and *S13*.
- Quantum arbitration allows two parties to reach an agreement on a value in a secure manner, this can be used to protect the fairness of signalling messages using QT: *S13*. Studies have been classified based on approach and exact area in which the study was conducted, table III.

Table 3: Selected Studies and Classifications

Studies (SID)	Approach
S4, S5, S8, S12, S13	Quantum Cryptography
S3, S7, S9	Quantum Teleportation
S1, S6, S10	Quantum Networks
S2, S11	Other areas Quantum Communication

RQ2: How do the various quantum-based signalling solutions found in answer to RQ1 differ in terms of technology, technique, approach, method, and limitations?

This question aimed to explore, identify and compare the commonly used quantum technology, techniques, methods, and protocols for secure signalling in OSI L1/L2. The analysis revealed that the application of quantum communication approaches, techniques, methods, and protocols has the potential to significantly enhance signalling security in these layers. However, a notable gap exists in terms of interoperability and standardization, hindering seamless interaction of these protocols within carrier networks. Many methods and protocols are still in the experimental stage and have not been field-tested, primarily due to their complexity, high cost, and hardware compatibility issues, limiting their adoption in current telecommunications. Some interesting findings include the discovery of quantum arbitration and quantum secure signature studies that contribute to QC. While most studies rely on QKD through Qcrypt, others combine aspects of QKD and QT. However, a lack of standardization for interoperability and interconnectivity poses challenges for IRN over IPX. Security concerns related to distance, technical challenges, attenuation, and the no-cloning theorem for quantum repeaters and communication networks were identified. Despite the promise of QT, it still faces challenges such as freespace losses and decoherence. One practical quantum technology suitable for wide area network (WAN) infrastructure is Photonic qubits, which spans various classifications, followed by optical modes. Additionally, a noteworthy development is the "Quantum noise-assisted coherent radio-over-fiber cipher system" that employs quantum noise to encrypt radio signals, providing protection for signalling messages transmitted over radio waves. Summary is outlined in table 5 below:

Table 4: Quantum-based signalling comparisons

Quantum Technology	Quantum Technique	SID#	Quantum Method/Protocol	Limitations
Photonic Qubits	Quantum Cryptography	S4, S5 S8, S12, S13 Related studies (S1, S2, S10, and S11)	Continuous-variable QKD, Discrete-variable QKD, Hybrid QKD (entanglement-based), CV-DV QKD which can be based on BB84, Ekert91, BM92, DQKD protocol.	*Security:* limited distance, technical challenges, and atmospheric attenuation. *Interoperability: L1/L2* different standards, and integration complexity. *Implementation:* High complexity, costs, environmental sensitivity, and low-key exchange
Photonic Qubits, Optical Modes, Trapped Ions, Superconducting Circuits	Quantum Teleportation	S3, S7, S9	QSDC protocol based on Hyper-entangled states, Two-step QDC scheme, DL04 protocol, CV-AQS, DQSDC, Q2BC	*Security:* Entropy, eavesdropping risk, losses: Free-Space, attenuation, decoherence. *Interoperability: L2* Standards mismatch, relies on complex quantum channel. *Implementation:* Detector efficiency for small photons, quantum channel noise, distance limits, regulations, distance limits, deployment costs.
Photonic qubits, Optical Modes	Quantum Networks	S1, S6, S10 Related studies (S2, S9, S11, and S12)	Quantum repeaters, Quantum Key Networks, Quantum internet, PQNI Entangled pairs, Hybrid CV-DV QCN, Hybrid QCN, Hybrid QKD (entanglement-based), QNP using Connection-Oriented entanglement distribution protocol.	*Security:* attack risk during node repeater decryption, distance limits, and technical issues, atmospheric attenuation, No-cloning theorem, quantum noise and distance. *Interoperability: L1/L2* Standardisation, compatibility for multiple protocols and hardware. *Implementation:* Lab development challenges, high costs, scalability issues, entanglement fragility, incomplete quantum repeater development.
Photonic qubits, Optical Modes	Other QC's	S2, S11	Quantum arbitration, Quantum signature	*Security:* technical challenges *Interoperability: L2* Standardisation, compatibility for multiple protocols *Implementation:* Complexity

In table 5 a greater attempt to identify most used quantum technology, quantum techniques, Method and protocol for secure signalling approaches also taking into an account the constraints while analysing the details

of preselected papers was undertaken. Our extensive analysis of the studies demonstrates that application of QC approaches with their various techniques, methods and protocols can significantly improve signalling security on L1/L2. However, there seems to be gap within interoperability and intercommunication for the protocols, there exists multiple approaches which have not been standardized for interoperability when protocols interact in carrier networks. Further investigation of the literature as categorised in RQ1 found that extensive research demonstrate that multiple methods and protocols are still being developed and tested in the laboratory with no deployment. We also found that the physical implementation for these approaches and techniques in the field are complex, costly, and not compatible with multiple hardware. This will make them not to be a best candidate for adoption in the current telecommunications domain. Interesting aspects of quantum arbitration and quantum secure signature were discovered in studies (Abulkasim et al., 2019) and (Li et al., 2019). These studies do not actually form part of Qcrypt nor QT however they are part of QC. Most studies rely on Qcrypt through QKD, while others mainly focused on QN combining both aspects of QKD and QT. These studies have one common factor lack of standardization for interoperability and interconnectivity which further widens the gap for IRN over IPX. Another aspect of security which was identified is the various levels of attack that may arise due to distance, technical challenges, attenuation and the no-cloning theorem for quantum repeaters and communication networks. QT though looks promising it still suffers from freespace losses and decoherence. The most practical quantum technology identified across wide area network (WAN) infrastructure has been Photonic qubits, that managed to cut across the four classifications followed by optical modes. The last interesting aspect of quantum signalling is the "Quantum noise-assisted coherent radio-over-fiber cipher system" which uses quantum noise to encrypt radio signals. This can be used to protect signalling messages that are transmitted over radio waves *S12, S13.*

RQ3: How strong is the evidence in support of the different quantum-based signalling security solutions?

Table 5: Strength of evidence

Method/ Protocol	Strengths	Constraints/Limitations
CV-QKD	High key rates, Compatible with existing telecom infrastructure, robust to noise and scalable.	Comple data processing, Less mature technology, Susceptible to certain attacks e.g., Photon splitting attack, requires high-quality components.
DV-QKD & DQKD	Proven security, tolerance to noise, efficient key generation, well-established technology. Efficient modulation scheme, less susceptible to noise and interference.	Single-photon detectors, less compatible with existing technology infrastructure, splitting attacks vulnerability, complex modulation schemes.
Hybrid CV-DV QKD	Larger number of entangled states, tolerant to noise, more secure and versatile. Serve as backbone for quantum Internet.	Less mature technology, more complex difficult to implement and deploy, high-quality components, less efficient to DV-QKD.
QSDC	Uses hyper entangled states, Bell states for secure transmission, secure information without key generation or sharing, present-day linear optics measurement, realized for practical applications, facilitate interconnection of QCN	Noisy quantum channels vulnerable, Relies on quantum memory, less mature technology, more complex, expensive equipment
DQSDC	Deterministic, unconditionally secure, high key rates and simple to implement	Less efficient than QSDC and BB84, complex, requires entanglement, vulnerable to noise
CV-AQS	Uses arbitrator and verifier, Signature cannot be denied or forged. Use dense coding and teleportation. Can be used on lossy and lossless channel, security by two-mode squeezed vacuum state teleportation. Scalable, compatible with existing telecommunication infrastructure.	Complex data processing, less mature technology, susceptible to photon number splitting attack, high-quality components
DL04	Discrete-variable (DV) Proven security, efficient and robust to noise, simple implement. Continuous-variable (CV) does not require entanglement, robust to noise, efficient and scalable	Requires entanglement, collective attack, difficult to scale. Not provably secure, gaussian attack, not fully implemented
Q2BC	BB84 protocol with proven security, robust to noise and simple to implement.	Requires entanglement, collective attack, difficult to scale. Not provably secure, gaussian attack, not fully implemented

Method/ Protocol	Strengths	Constraints/Limitations
PQNI	Supports quantum packet communication, robust to noise, scalable and secure against eavesdropping	Requires entanglement, complex and expensive, not yet fully developed, denial of service attacks
Quantum repeaters	data plane network using virtual circuits on link layer protocol. Scalability and flexibility, resilient to noise and extends distance of QC	limited implementation for control plane protocol. Requires entanglement, denial of service attacks, not fully developed, complex and expensive.

In Table 6, an important consideration is the choice of QC method/protocol. Which method is most appropriate depends on several factors, including the required key rates, the distance between nodes and the level of security required. Promising methods/protocols include CV-QKD, Hybrid CV-DV QKD, CV-AQS, QSDC and quantum repeaters. Another important consideration is the interoperability of the chosen method/protocol with the existing telecommunications infrastructure. This is particularly important in the context of IRN, which often involves multiple operators and a variety of systems and networks. The chosen method/protocol must also be able to operate at both L1/L2 (i.e. the physical and data link layers) to send and receive raw data over the physical network infrastructure.

RQ4: What are the potential implications of these findings of the research when developing a new quantum-based signalling security solution for NGN-enabled IoT over IPX?

This question focused on the potential implications of QC methods and protocols for the development of a new quantum-based signalling security solution for IRN over IPX. The choice of method/protocol for this application depends on key factors such as required key rates, node distance, and security levels. Strengths to consider include compatibility with existing infrastructure, interference robustness, scalability, and security. Limitations involve complexity, cost, and technology maturity. The table analysis suggests several suitable methods/protocols. CV-QKD is compatible with existing infrastructure and robust but less mature. Hybrid CV-DV QKD offers enhanced security but is complex and less mature. QSDC is highly secure but costly and complex. CV-AQS is scalable and compatible but less mature. Quantum repeaters could extend range but are in early development stages. The choice depends on specific network requirements, making a careful evaluation of strengths and limitations is essential. Given the complexity and multiple operators in IRN networks, interoperability across systems in OSI L1/L2 is crucial. The selected method/protocol should operate on both layers to handle raw data over the physical network. Developing a quantum-based signaling security model for IRN over IPX at OSI L1/L2 is challenging but provides unprecedented level of security. In summary the study contributes following insights:

- Quantum-based signalling security approaches, techniques, methods/protocols studied, compared, and presented.
- The advantages of quantum-based signalling techniques are highlighted.
- The quantum-based technology strengths and limitations are highlighted.
- Performance and security based on distance for different QC methods/protocols in the context of IRN are presented.

Lastly, signalling security gaps are identified, and future research directions for enhanced protection of IRN over IPX are discussed.

6. Threats to Validity

Conducting a SLR has risks, especially in rapidly evolving fields like quantum-based secure signalling. We mitigated these risks by defining precise search criteria and using logical operators to broaden our search scope across identified impactful databases. We focused on articles from 2018 to 2023, ensuring our review represents the current state of quantum-based secure signalling for IRN over IPX at OSI L1/L2 covering NB-IoT. While the risk of missing unpublished papers exists, we believe our chosen databases are comprehensive and respected in the field and our RQ presents current knowledge and recent developments in QC. Our rigorous approach increases the validity and reliability of our review, we believe we took reasonable steps in minimizing associated risks.

7. Conclusion

In this review, we focused on quantum-based secure signalling solutions for NGN-enabled NB-IoT traffic, evaluating their approaches, methods, and limitations. Our comprehensive analysis involved 13 studies that met our selection criteria. These quantum-based solutions are still in their infancy, therefore not ready for widespread use. They leverage QC techniques and protocols to enhance signalling security at the OSI L1/L2 layers, offering protection against emerging threats like quantum cryptanalysis, HiddenArt, and HNDTL. Nevertheless, they come with challenges. Issues related to interoperability and intercommunication can leave networks vulnerable to side-channel and physical attacks. High implementation costs and compatibility concerns with existing hardware restrict their adoption. Additionally, the lack of standardized protocols for interoperability across operator networks hinders broader adoption. Signalling protocols play a crucial role in ensuring seamless interconnection, particularly for securing roaming signals and subscribers in the context of IRN over IPX, which has become increasingly vital for global infrastructure. Developing quantum-based security solutions requires a unique blend of expertise, encompassing roaming experience, cyber-security skills, and quantum computing knowledge. This is essential to create a model that can secure the IPX-IRN ecosystem for NGN networks like 5G. While quantum technologies have been extensively studied across various fields, a significant gap exists in developing quantum models tailored for the integrated and standardized nature of signalling interconnect and roaming L1/L2 networks over IPX to secure roaming traffic. Our study is one of the initial efforts to address this gap by systematically reviewing quantum approaches, methods, and protocols to create a quantum-based signalling model for IRN over IPX, particularly for NB-IoT traffic. This is crucial, as these networks underpin critical infrastructure, including power grids, health systems, and financial markets, making them susceptible to unconventional surveillance threats.

8. Future Directions

Future research should focus on evaluating the performance and security of QC methods in IRN, customizing methods for IRN needs, creating interoperable models for seamless protocol interaction, and standardizing quantum protocol operation on OSI L1/L2. Developing a secure quantum-based signalling model for IRN to enhance OSI L1/L2 signalling security over IPX for NB-IoT.

References

Abdel Hakeem, S.A., Hussein, H.H. & Kim, H. 2022. Security Requirements and Challenges of 6G Technologies and Applications. *Sensors*, 22(5).

Abulkasim, H., Alsuqaih, H.N., Hamdan, W.F., Hamad, S., Farouk, A., Mashatan, A. & Ghose, S. 2019. Improved Dynamic Multi-Party Quantum Private Comparison for Next-Generation Mobile Network. *IEEE Access*, 7: 17917–17926.

COTS Staff. 2023. Enea Urges EU PEGA Committee: Broaden the focus beyond spyware to combat mobile surveillance threats and signaling infrastructure exploitation. – COTS Journal. https://www.cotsjournalonline.com/index.php/2023/05/16/enea-urges-eu-pega-committee-broaden-the-focus-beyond-spyware-to-combat-mobile-surveillance-threats-and-signaling-infrastructure-exploitation/?utm_campaign=Signalling%20Firewall&utm_medium=email&_hsmi=259452321&_hsenc=p2ANqtz-_B7VNxlP2rskn0onouuyW6vgqdGKilYLEcaGJwnJRtXTSePiyPM7Vf8mGbwlh4JHGKvwmchYq11Qzs7hKLqozlMxCx1g&utm_content=259452321&utm_source=hs_email 23 May 2023.

Asfaw, A., Blais, A., Brown, K.R., Candelaria, J., Cantwell, C., Carr, L.D., Combes, J., Debroy, D.M., Donohue, J.M., Economou, S.E., Edwards, E., Fox, M.F.J., Girvin, S.M., Ho, A., Hurst, H.M., Jacob, Z., Johnson, B.R., Johnston-Halperin, E., Joynt, R., Kapit, E., Klein-Seetharaman, J., Laforest, M., Lewandowski, H.J., Lynn, T.W., McRae, C.R.H., Merzbacher, C., Michalakis, S., Narang, P., Oliver, W.D., Palsberg, J., Pappas, D.P., Raymer, M.G., Reilly, D.J., Saffman, M., Searles, T.A., Shapiro, J.H. & Singh, C. 2022. Building a Quantum Engineering Undergraduate Program. *IEEE Transactions on Education*.

Baraković Husić, J., Bajrić, H. & Baraković, S. 2012. Evolution of Signaling Information Transmission. *ISRN Communications and Networking*, 2012: 1–9.

Chen, N., Shuai, S., Yan, B., Xu, N. & Pei, C. 2020. Multi-hop quantum communication based on greenberger-horne-zeilinger states and teleportation. *IEEE Access*, 8: 52052–52061.

Chih-Lin, I., Han, S., Xu, Z., Sun, Q. & Pan, Z. 2016. 5G: rethink mobile communications for 2020+. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2062): 20140432. https://royalsocietypublishing.org/doi/10.1098/rsta.2014.0432.

Cho, J.Y. 2019. Securing optical networks by modern cryptographic techniques. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11875 LNCS: 120–133. https://link.springer.com/chapter/10.1007/978-3-030-35055-0_8 23 August 2023.

Djordjevic, I.B. 2022. Physical-Layer Security, Quantum Key Distribution, and Post-Quantum Cryptography. *Entropy 2022, Vol. 24, Page 935*, 24(7): 935. https://www.mdpi.com/1099-4300/24/7/935/htm 25 March 2023.

Dutta, A. 2022. 5G and Beyond Security Challenges and 5G & Beyond : Security Perspective. *Johns Hopskins University*. https://nsf-nextg-security.cs.ucsb.edu/sites/default/files/2020-11/Ashutosh Dutta.pdf.

ENISA. 2018. Signalling Security in Telecom SS7/Diameter/5G. *European Union Agency for Network and Information Security*, (March): 1–30. www.enisa.europa.eu.

Garcia Cid, M.I., Álvaro González, J., Ortíz Martín, L. & Del Río Gómez, D. 2022. Disruptive Quantum Safe Technologies. *ACM International Conference Proceeding Series*. https://dl.acm.org/doi/10.1145/3538969.3544484 23 August 2023.

Garzon, S.R., Yildiz, H. & Küpper, A. 2022. Towards Decentralized Identity Management in Multi-stakeholder 6G Networks. http://arxiv.org/abs/2203.00300.

GSM Association. 2021. *5GS Roaming Guidelines*. https://www.gsma.com/newsroom/wp-content/uploads//NG.113-v2.0.pdf 19 August 2021.

GSMA. 2020. *GSM Association Non-confidential Official Document IR.67-DNS Guidelines for Service Providers and GRX and IPX Providers DNS Guidelines for Service Providers and GRX and IPX Providers Copyright Notice Antitrust Notice GSM Association Non-confidential Offic*.

Holtmanns, S. 2016. Interconnection Security Standards - We Are All Connected. *Journal of ICT Standardization*, 4(1): 1–18. https://riverpublishers.com/journal_read_html_article.php?j=JICTS/4/1/1 27 November 2021.

Holtrup, G., Lacube, W., Dimitri, ;, David, P., Mermoud, A., Gérôme Bovet, ; & Lenders, V. 2021. 5G System Security Analysis.

Intelligence, K. & IBASIS. 2020. *5G ROAMING: MNO REQUIREMENTS & OPPORTUNITIES*. https://ibasis.com/wp-content/uploads/2020/05/5G-Roaming_-_IPX_Requirements__Opportunities.pdf.

ITU-R. 2015. IMT Vision-Framework and overall objectives of the future development of IMT for 2020 and beyond M Series Mobile, radiodetermination, amateur and related satellite services. *Recommendation ITU-R M.2083-0*, (September 2015): 1–21. http://www.itu.int/ITU-R/go/patents/en 11 August 2022.

ITU-T. 2021. ITU Workshop on 'Improving the security of signalling protocols'. https://www.itu.int/en/ITU-T/Workshops-and-Seminars/2021/1129/Pages/default.aspx 17 September 2022.

Keller, R., Castellanos, D., Sander, A., Robison, A. & Abtin, A. 2021. THE 5G SYSTEM ROAMING ARCHITECTURE. *Review*: 1–36. https://www.ericsson.com/4981f6/assets/local/reports-papers/ericsson-technology-review/docs/2021/roaming-in-the-5g-system.pdf 24 October 2021.

Khawasik, M., Elsayed, W., Rashad, M. & Younes, A. 2022. A Secured Quantum Two-Bit Commitment Protocol for Communication Systems. *IEEE Access*, 10: 50218–50226.

Kim, H. 2020. 5G core network security issues and attack classification from network protocol perspective. *Journal of Internet Services and Information Security*, 10(2): 1–15.

Køien, G.M. 2021. On Threats to the 5G Service Based Architecture. *Wireless Personal Communications*, 119(1): 97–116. https://doi.org/10.1007/s11277-021-08200-0 27 August 2021.

Kozlowski, W., Dahlberg, A. & Wehner, S. 2020. Design-ing a Quantum Network Protocol. *Proceedings of the 16th International Conference on emerging Networking EXperiments and Technologies*, 16. https://doi.org/10.1145/3386367.3431293 26 March 2023.

Li, J., Jia, Q., Xue, K., Wei, D.S.L. & Yu, N. 2022. A Connection-Oriented Entanglement Distribution Design in Quantum Networks. *IEEE Transactions on Quantum Engineering*, 3.

Li, J., Zhou, Z., Wang, N., Tian, Y., Yang, Y.G. & Zheng, Y. 2019. Deterministic Quantum Secure Direct Communication Protocol Based on Hyper-Entangled State. *IEEE Access*, 7: 43948–43955.

Lou, X., Long, H., Tang, W., Yang, Y. & Li, J. 2019. Continuous-Variable Arbitrated Quantum Signature Based on Dense Coding and Teleportation. *IEEE Access*, 7: 85719–85726.

Lutu, A., Perino, D., Bagnulo, M. & Bustamante, F.E. 2021. Insights from operating an IP exchange provider. *SIGCOMM 2021 - Proceedings of the ACM SIGCOMM 2021 Conference*: 718–730.

Lutu, A., Trevisan, M., Safari Khatouni, A., Mandalari, A.M., Custura, A., Mellia, M., Alay, O., Bagnulo, M., Bajpai, V., Brunstrom, A., Ott, J. & Fairhurst, G. 2021. Measuring Roaming in Europe: Infrastructure and Implications on Users QoE. *IEEE Transactions on Mobile Computing*.

Manzalini, A. 2020. Quantum communications in future networks and services. *Quantum Reports*, 2(1): 221–232.

Moher, D., Liberati, Alessandro, Tetzlaff, J., Altman, Douglas G, Liberati, A & Altman, D G. 2009. Reprint-Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. http://www.annals.org/cgi/content/full/151/4/264. 12 August 2023.

Ngobeni, T., Kabaso, B. & Mukherjee, A. 2020. A generic framework for the implementation of a secure quantum teleportation infrastructure. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems, icABCD 2020 - Proceedings*.

Ntanos, A., Lyras, N.K., Zavitsanos, D., Giannoulis, G., Panagopoulos, A.D. & Avramopoulos, H. 2021. Leo satellites constellation-to-ground qkd links: Greek quantum communication infrastructure paradigm. *Photonics*, 8(12).

Okoli, C. 2015. A guide to conducting a standalone systematic literature review. *Communications of the Association for Information Systems*, 37(1): 879–910.

Pan, D., Li, K., Ruan, D., Ng, S.X. & Hanzo, L. 2020. Single-Photon-Memory Two-Step Quantum Secure Direct Communication Relying on Einstein-Podolsky-Rosen Pairs. *IEEE Access*, 8: 121146–121161.

Pirandola, S., Eisert, J., Weedbrook, C., Furusawa, A. & Braunstein, S.L. 2015. Advances in Quantum Teleportation. *Nature Photonics*, 9(May): 641–652. http://arxiv.org/abs/1505.07831%5Cnhttp://www.arxiv.org/pdf/1505.07831.pdf.

Suomalainen, J., Kotelba, A., Kreku, J. & Lehtonen, S. 2018. Evaluating the efficiency of physical and cryptographic security solutions for quantum immune iot. *Cryptography*, 2(1): 1–20.

Tan, Z., Ding, B., Zhang, Z., Li, Q., Guo, Y. & Lu, S. 2021. Device-Centric Detection and Mitigation of Diameter Signaling Attacks against Mobile Core. In *2021 IEEE Conference on Communications and Network Security (CNS)*. IEEE. https://ieeexplore-ieee-org.ezproxy.cput.ac.za/document/9705031/ 15 August 2022.

Tan, Z., Ding, B., Zhao, J., Guo, Y. & Lu, S. 2022. Breaking Cellular IoT with Forged Data-Plane Signaling: Attacks and Countermeasure. https://doi.org/10.1145/3534124 15 August 2022.

Tanizawa, K. & Futami, F. 2022. IF-Over-Fiber Transmission of OFDM Quantum-Noise Randomized PSK Cipher for Physical Layer Encryption of Wireless Signals. *Journal of Lightwave Technology*, 40(6): 1698–1704.

Tanizawa, K. & Futami, F. 2020. Quantum Noise-Assisted Coherent Radio-Over-Fiber Cipher System for Secure Optical Fronthaul and Microwave Wireless Links. *Journal of Lightwave Technology*, 38(16): 4244–4249.

Tezergil, B. & Onur, E. 2021. Wireless Backhaul in 5G and Beyond: Issues, Challenges and Opportunities.

United Nations. 2016. The promotion, protection and enjoyment of human rights on the Internet. *Online Document*, 10802(December 2015): 1–4. https://www.article19.org/data/files/Internet_Statement_Adopted.pdf 29 August 2022.

Wang, M., Zhu, T., Zhang, T., Zhang, J., Yu, S. & Zhou, W. 2020. Security and privacy in 6G networks: New areas and new challenges. *Digital Communications and Networks*, 6(3): 281–291. https://doi.org/10.1016/j.dcan.2020.07.003.

You, X., Wang, C.X., Huang, J., Gao, X., Zhang, Z., Wang, M., Huang, Y., Zhang, C., Jiang, Y., Wang, J.J., Zhu, M., Sheng, B., Wang, D., Pan, Z., Zhu, P.P., Yang, Y., Liu, Z., Zhang, P., Tao, X., Li, S., Chen, Z.Z., Ma, X., Chih-Lin, I., Han, S., Li, K., Pan, C., Zheng, Z., Hanzo, L., Shen, X.S., Guo, Y.J., Ding, Z., Haas, H., Tong, W., Zhu, P.P., Yang, G., Wang, J.J., Larsson, E.G., Ngo, H.Q., Hong, W., Wang, H., Hou, D., Chen, J., Chen, Z.Z., Hao, Z., Li, G.Y., Tafazolli, R., Gao, Y., Poor, H.V., Fettweis, G.P. & Liang, Y.C. 2021. *Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts*. Science in China Press. https://www.researchgate.net/publication/347799507_Towards_6G_wireless_communication_networks_vision_enabling_technologies_and_new_paradigm_shifts 9 July 2021.

Yu, C., Chen, S., Wang, F. & Wei, Z. 2021. Improving 4G/5G air interface security: A survey of existing attacks on different LTE layers. *Computer Networks*, 201: 108532.

Yu, N., Lai, C.Y. & Zhou, L. 2021. Protocols for Packet Quantum Network Intercommunication. *IEEE Transactions on Quantum Engineering*, 2.

Zimmer, P., Weinreich, R., Zenger, C.T., Sezgin, A. & Paar, C. 2021. Keys from the Sky: A First Exploration of Physical-Layer Security Using Satellite Links. *IEEE International Conference on Communications*.

Integrating Enterprise Architecture into Cybersecurity Risk Management in Higher Education

Mafika Nkambule[1], Joey Jansen van Vuuren[1] and Louise Leenen[2]
[1]Tshwane University of Technology, Pretoria, South Africa
[2]University of the Western Cape and CAIR, Cape Town, South Africa

nkambulemw@tut.ac.za
Jansenvanvuurena1@tut.ac.za
lleenen@uwc.ac.za

Abstract: Cybercriminals constantly seek new methods to infiltrate a company's defences, making cybersecurity investments essential. Enterprise architecture (EA) provides a systematic risk detection and mitigation process by emphasising the interdependencies between systems, data, processes, people, and other factors. This paper provides a comprehensive approach, also referred to as a process, based on EA to assist African universities in developing a comprehensive cybersecurity plan. The EA process comprises four pillars: business architecture, data architecture, application architecture, and technology architecture. African universities can develop a comprehensive cybersecurity strategy using an EA approach in cybersecurity to achieve institutional goals and objectives. The potential attack surface comprises isolated EA components and their interconnections. This article comprehensively examines various EA processes such as business, information, application, and technology architecture. These processes are carefully analysed to evaluate the organisational structures and uncover opportunities to enhance security protocols. Additionally, we delve deep into abstract security patterns, seeking to cultivate an environment of trustworthiness within complex systems. Our research findings underscore the significant potential within African higher education institutions. By embracing a model-based approach to risk analysis and mitigation, these institutions can fortify their cybersecurity defences and bolster their capabilities to ensure uninterrupted business operations and enhance overall resilience in the face of evolving security challenges. When we combine EA and information security (ICS), we uncover many vulnerabilities malicious actors might exploit. By embracing a holistic EA-based methodology, institutions can craft and implement robust security protocols to safeguard their components and connections. Leveraging EA, our proposed integrated approach aims to forge a comprehensive cybersecurity risk management strategy tailored to the African higher education sector. This strategy seeks to facilitate the identification of critical elements and their intricate interrelationships, thus formulating an effective defence strategy against potential cyber threats. The synergy between EA and cybersecurity within African universities promises to elevate cybersecurity practices, ensure uninterrupted business operations, and fortify the continent's resilience.

Keywords: Enterprise architecture, Cybersecurity risk management, Higher education, Integration, Framework

1. Introduction

1.1 Background

Cyber threats have evolved in complexity and frequency in today's interconnected digital landscape. With their vast resources, universities have become enticing targets for cybercriminals, who covet assets like trademarks, student data, and collaborative research ventures (Lallie et al., 2023). Within the academic realm, cybersecurity presents a multifaceted challenge, as it demands the perpetual preservation of digital resources' confidentiality, integrity, and accessibility, all while contending with limited resources and a diverse user base.

Despite notable advancements in cybersecurity, a gap persists between the prevailing security needs and the broader institutional framework of our modern world. It has become increasingly imperative for higher education institutions to align their cybersecurity initiatives with their overarching goals, systems, and procedures (Dlamini et al., 2011). By establishing this alignment, security processes can seamlessly integrate into an organisation's operational framework, taking on a proactive role.

EA, a comprehensive approach to managing a corporation Maulana, Azmi and Phon (2023), may hold the key to bridging the existing gap in cybersecurity handling. EA provides a holistic view of an organisation, encompassing its processes, procedures, data flows, and technological infrastructure. Institutions seeking to enhance their risk management capabilities and adapt to changing circumstances would be well-advised to merge their EA and cybersecurity strategies. This integration can lead to more robust and proactive security measures in an ever-evolving digital landscape.

1.2 Purpose and Relevance of the Study

The central aim of this research is to delve into the prospective advantages associated with the incorporation of EA processes into a comprehensive cybersecurity risk management framework tailored specifically for higher education institutions. Given the alarming surge in cybersecurity attacks and the emergence of formidable vulnerabilities within the realm of higher learning institutions—occasionally driven by personal interests and gains, as outlined by Lallie et al. (2023), there exists an urgent imperative to formulate a strategy. This strategy must not only effectively tackle the operational challenges posed by cybersecurity threats but also harmoniously align with the holistic academic and research objectives of these institutions, from start to finish.

This research work is guided by the following pivotal goals:

- **Deep Dive Analysis:** We aim to assess the current cybersecurity frameworks, and pinpoint both shortcomings and avenues for enhancement.
- **EA and Cybersecurity Synergy:** The study seeks to understand how integrating EA can act as a bedrock in formulating a malleable yet steadfast cybersecurity risk management framework. This encompasses assessing the convergence of institutional objectives, technological underpinnings, data pathways, and protective mechanisms.
- **Blueprint Creation:** Our ambition is to curate an actionable guideline based on EA principles, crafted for academic institutions. This blueprint is envisioned to equip institutions with a strategic trajectory to amplify their digital protection, ensure operational consistency, and match protective endeavours with the broader institutional vision.

In pursuing these objectives, the study hopes to contribute significantly to the discourse on cybersecurity in higher education, offering institutions a nuanced strategy that goes beyond traditional, siloed approaches to cyber risk management. By fostering a deeper integration of EA and cybersecurity, this research aspires to pave the way for a more secure, resilient, and adaptive digital environment within academia.

2. Literature Review

2.1 Cybersecurity in Higher Education

Institutions of higher learning are increasingly becoming main targets for cyber criminals due to their huge repositories of student data, intellectual property, and collaborative research projects (Nasir et al., 2023). There is always that huge obligation to ensure the confidentiality, integrity, and availability of digital resources, paired with budget constraints and a diverse user environment. This makes the task of cybersecurity in the academic environments an extremely difficult undertaking. According to Eltahir and Ahmed (2023), the move for institutions of higher learning to go online has made them even more vulnerable to cyber criminals.

2.2 The Role of Enterprise Architecture in Cybersecurity

There are many definitions of enterprise architecture, but this study adopted the definition by the Pereira and Sousa (2005), who defines EA as the process by which organizations standardize and organize IT infrastructure to align with business goals. EA, by design, looks at all aspects of an organisation, and these various aspects are what attackers use to strike. An EA based approach, therefore, offers a promising approach to bridging the gap between cybersecurity measures and the overall institutional architecture.

EA lays out a structured and comprehensive blueprint of an organisation's operations, encompassing its goals, processes, information flows, and technological infrastructure (Kotusev et al., 2023). By making sure EA is integrated with cybersecurity measures, there is an opportunity to create a unified, resilient, and adaptive security architecture that is responsive to threats and aligns with the institution's goals and changing landscape.

2.3 Existing Frameworks and Their Limitations

It is important to note that numerous cybersecurity risk management frameworks are used across all industries, including higher eductaion. While each of these frameworks offers its own set of guidelines and practices, they often fall short in synchronizing seamlessly with the comprehensive objectives, infrastructure, and day-to-day workings of organisations. Many of these models typically react to issues post-occurrence, rather than embedding themselves organically into the core strategies of institutions.

While there is undoubtedly a plethora of frameworks available, we first present a unique overview of the various cybersecurity frameworks that are most widely recognized and utilized across different industries as identified an documented by (Taherdoost, 2022). For a clearer comparison, we have encapsulated our observations in the table below, contrasting these mainstream frameworks against our innovative approach. Through the integration of EA with conventional cybersecurity strategies, there is a potential for a more harmonized, robust, and agile security structure, one that not only anticipates threats but also aligns perfectly with an institution's evolving objectives and landscape.

Table 1: Cybersecurity Frameworks Comparison

#	SOURCE	EXISTING FRAMEWORK	COMPARING EXISTING FRAMEWORKS WITH OUR PROPOSED EA-BASED APPROACH TO DEVELOP A CYBERSECURITY RISK MANAGEMENT FRAMEWORK
1	(White & Sjelin, 2022)	**NIST Cybersecurity Framework (CSF)**	While the NIST CSF provides a detailed set of cybersecurity practices, its primary focus is on managing and reducing cybersecurity risk. Using the EA approach, with the comprehensive overview of business, data, application, and technology, inherently ties cybersecurity to the organization's strategic objectives.
2	(Humphreys, 2016)	**ISO/IEC 27001**	ISO/IEC 27001 is largely about defining controls based on a risk assessment. EA, on the other hand, allows for a more holistic risk assessment that looks at risks from multiple dimensions - business processes, data flows, and applications.
3	(Groš, 2021)	**CIS Critical Security Controls**	CIS controls are prescriptive and are designed to provide a set of actions for cyber defence. In contrast, the EA approach is inherently customizable, designed to adapt to the unique needs and structures of individual organizations.
4	(IGNAT, 2022)	**FAIR (Factor Analysis of Information Risk)**	FAIR is primarily a quantitative risk analysis model, focusing heavily on the financial implications of cybersecurity risks. The EA approach offers a broader view that aligns cybersecurity with business goals and objectives.
5	(Ahmed et al., 2022)	**MITRE ATT&CK**	Focuses primarily on the tactics, techniques, and procedures used by adversaries against organizations. While it's a comprehensive knowledge base of adversary behaviour, it doesn't inherently provide an overall organizational strategy., which the EA approach offers.
6	(Morse & Raval, 2008)	**PCI DSS (Payment Card Industry Data Security Standard)**	PCI DSS is specific to payment card security. While it's crucial for companies dealing with card payments, it's narrow in scope. In contrast, an EA-based approach offers a more holistic view, covering all business, data, application, and technology domains.
7	(De Haes et al., 2013)	**COBIT (Control Objectives for Information and Related Technologies)**	Both COBIT and EA offer governance structures, but the EA-based approach has a core focus on aligning cybersecurity strategies with business objectives.
8	(Burkett, 2012)	**SABSA (Sherwood Applied Business Security Architecture)**	While SABSA is an enterprise security architecture framework, it heavily focuses on security. The EA-based approach ensures a balanced representation of business, data, application, and technology domains, emphasizing their co-dependency.
9	(Sharkov, 2020)	**C2M2 (Cybersecurity Capability Maturity Model)**	C2M2 is designed to measure the maturity of an organisation's cybersecurity capabilities. An EA-based approach, while valuing maturity measurements, also emphasises the holistic alignment of all cybersecurity measures with the business's overarching goals.

While each of the listed frameworks provides important guidelines, tools, and metrics in specific areas of cybersecurity and governance, the proposed EA-based approach stands out due to its holistic perspective, adaptability, and emphasis on aligning cybersecurity with overarching organizational goals. It provides a comprehensive process that bridges the multifaceted dimensions of an organization, making it a potentially valuable addition to the cybersecurity landscape.

3. The Importance of Cybersecurity Risk Management in African Higher Education Institutions

3.1 Current Cyber Threat Landscape

While it is acknowledged that cybersecurity challenges are a global phenomenon affecting institutions worldwide, this study specifically narrows its focus to the African context. In particular, African higher education institutions face a growing threat from cyber adversaries due to their increasing reliance on digital technologies (Eltahir & Ahmed, 2023). The threat landscape constantly evolves, with cybercriminals using sophisticated techniques to gain unauthorised access to sensitive data. The most common types of cyber threats include phishing attacks, malware, ransomware, and denial-of-service attacks (Rana & Sharma, 2023). These threats can result in data breaches, financial losses, and reputational damage to institutions.

3.2 Unique Challenges Faced by African Institutions

As indicated in section 3.1 above, this is a challenge for institutions across the globe. To ensure the effectiveness of the proposed framework, we narrowed our scope to focus only on challenges prevalent to African universities. According to Rana and Sharma (2023), African higher education institutions face unique challenges in implementing effective cybersecurity measures. These challenges include limited financial resources, inadequate cybersecurity expertise, and a lack of awareness among users. The diverse user environment, including students, faculty, and staff, makes implementing a unified cybersecurity strategy difficult (Kayode-Ajala, 2023). The lack of a comprehensive legal framework for cybersecurity in many African countries also poses a challenge (Bouke et al., 2023).

To address these challenges, it is essential for African higher education institutions to prioritise cybersecurity and invest in robust cybersecurity measures (Kayode-Ajala, 2023). It is for this reason that the implementation of an EA-based approach is proposed, to align cybersecurity measures with academic institutions' holistic objectives, infrastructure, and operations. By integrating EA with cybersecurity measures, institutions can create a unified, resilient, and adaptive security architecture that is responsive to threats and aligns with the institution's goals and changing landscape.

4. Understanding Enterprise Architecture

4.1 Definition and Components of EA

EA, in its essence, represents a meticulously detailed and all-encompassing roadmap of an organization's operations, entailing not just its objectives and processes but also the intricate flow of information and the underpinning technological infrastructure (Rozas et al., 2020). The beauty of EA lies in its ability to paint a panoramic portrait of an organization, granting stakeholders the insight to decipher the intricate symphony of components working in tandem to fulfil its ambitions. In concurrence with The Open Architecture Group's (TOGAF) insights, the facets of EA encompass a wide spectrum of elements (Maulana, Azmi, & Arshah, 2023) as elegantly illustrated in the visual representation provided in Figure 1 below:

- Business Architecture: Describes the organisation's business strategy, goals, processes, and organisational structure.
- Information Architecture: Describes the organisation's information assets, data flows, and information systems (Taherdoost, 2022).
- Application Architecture: Describes the organisation's software applications, interdependencies, and alignment with business objectives.
- Technology Architecture: Describes the organisation's hardware, network, and infrastructure components.

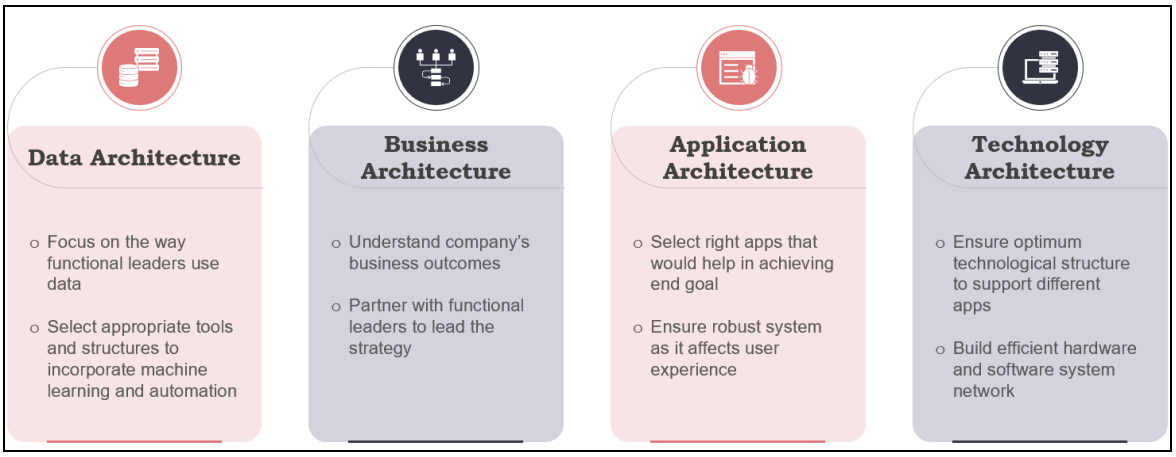


Figure 1: Enterprise Architecture Domains

5. Proposed EA-Based Cybersecurity Approach for Higher Education

For the distinctive challenges African academic institutions face in implementing robust cybersecurity controls, we propose a cybersecurity blueprint based on EA processes. This blueprint is specially proposed for the academic institutions, aspiring to synchronize cybersecurity initiatives with the encompassing goals, foundation, and functioning of these educational bodies in Africa and beyond. Figure 2 below depicts the elements to be covered within each domain of EA in relation to cybersecurity. The suggested blueprint encompasses these core elements:

5.1 EA Approach: Adopting an Organisational Architecture within Security Governance

Within the dynamic landscape of academic institutions in Africa, the fusion of organizational architecture with the realm of security governance takes on paramount significance. At its core, this fusion entails the meticulous crafting of a precisely tailored governance framework, purpose-built for the astute management of cybersecurity risks within the unique confines of higher education environments, as eloquently articulated by (Gloria Appiah, 2020).

- **Role Definition and Allocation:** A crucial initial step is to demarcate clear roles within the institution's cybersecurity landscape. By understanding who is responsible for what, institutions can foster both accountability and clarity. From frontline IT staff to top-tier management, each stakeholder's role in the cybersecurity matrix needs clear delineation.
- **Policy and Procedural Process:** Once roles are defined, the next step is to develop a comprehensive set of policies and procedures. These aren't just rulebooks but are dynamic guides that evolve with the changing cyber threat landscape. They act as the institution's first line of defence, guiding actions and setting the standards for cybersecurity measures.
- **Regulatory Adherence and Compliance:** Higher education institutions often deal with a plethora of sensitive data, making regulatory compliance non-negotiable. This facet ensures that the institution isn't just following internal guidelines but is also in line with regional, national, or international cybersecurity regulations, as applicable.
- **Risk Management Aligned with Institutional Vision:** The beauty of a tailored risk management strategy lies in its alignment with the broader goals of the institution. It's not just about countering threats but doing so in a way that seamlessly blends with the institution's educational, research, and administrative objectives. This ensures that while risks are mitigated, the institution's primary functions and aspirations aren't hampered.

Fundamentally, this element revolves around the conception of a resilient and adaptable framework, one in which cybersecurity seamlessly weaves itself into the very fabric of the institution, transcending the status of an afterthought.

5.2 EA Approach: Using Business Architecture for Security Strategy

At the heart of each forward-thinking educational establishment lies the essential quest to harmonize its business ambitions with the domain of cybersecurity pursuits. In this paper, we embark on a deep and

enlightening odyssey through this intricate interweaving, as thoughtfully illuminated by the discerning perspectives (Choudhuri et al., 2023).

- **Asset and Data Prioritization:** Recognizing and ranking the institution's pivotal assets and data repositories is a foundational step. It's about discerning what's at the core of the institution's operations and what would have the most profound impact if compromised.
- **Strategising with Vision:** Rather than adopting a one-size-fits-all approach, crafting a security plan that mirrors the institution's broader aspirations is crucial. Such alignment ensures that security measures don't just shield but also bolster the institution's growth trajectory.
- **Integration into Business Processes:** Cybersecurity is not an isolated domain but rather a pervasive one. It's essential to weave these protective measures seamlessly into the institution's everyday business activities, ensuring continuity and safety without hampering functionality.

In summary, this segment underscores the necessity of a holistic approach where cybersecurity is not an appendage but is integrally rooted in the institution's business architecture, promoting both growth and safety.

5.3 EA Approach: Using Information Architecture for Data Security

Within the expansive realm of academic institutions, safeguarding information assets and data stands paramount. This section delves into the nuances of this integration, ensuring a seamless blend of structure and protection.

- **Policy Formulation and Classification:** Before diving into protective measures, it's essential to have a structured approach to data. Creating clear policies and procedures aids in classifying data based on its sensitivity and relevance, ensuring each data subset gets its due attention.
- **Access Regulation and Retention:** Not all data should be accessible to everyone. Establishing rigorous access controls ensures that only authorized personnel can access specific data sets. Moreover, data retention policies help in determining the lifespan of data, ensuring obsolete or non-essential information doesn't clog the system or pose unnecessary risks.
- **Technical Fortifications:** While policies create a protective process, technical tools are the actual shields. By deploying state-of-the-art encryption protocols, robust firewalls, and proactive intrusion detection systems, institutions can defend against the ever-evolving threats of data breaches.

In essence, this segment emphasizes a layered approach to data protection. By intertwining robust information architecture with cutting-edge security protocols, institutions can ensure their data remains both organized and impenetrable.

5.4 EA Approach: Using Application Architecture for Software Security

In the digital fabric of modern educational institutions, the security of software applications holds a prominent position. This section illuminates the intricate balance between application structure and its subsequent protective layers.

- **Guided Software Lifecycle:** Every software application in the institution's arsenal undergoes a lifecycle from inception to decommission. Crafting meticulous policies and procedures ensures that each phase, be it development, testing, or deployment, is executed under a lens of security. This proactive approach ensures potential pitfalls are addressed long before they become threats.
- **Upholding Software Integrity:** It's not just about creating software; it's about creating software that's robust and resistant. Adopting secure coding practices right from the get-go ensures that the software's foundation is resilient. Regular vulnerability scans then ensure that this foundation remains uncompromised, catching potential weaknesses before they're exploited.
- **Proactive Defence Mechanisms:** Beyond just looking for weaknesses, penetration testing takes a more aggressive approach, simulating cyberattacks to test the software's defences. This "offensive defence" strategy gives institutions a clearer picture of potential real-world threats.

In summary, this segment champions a multi-faceted approach to software security. By harmonizing well-defined application architecture with advanced security methodologies, institutions can rest assured that their software arsenal is both functional and fortress-like.

5.5 EA Approach: Using Technology Architecture for Infrastructure Security

As academic institutions increasingly intertwine with the digital realm, ensuring the safety of their technological backbone becomes paramount. This section shines a spotlight on the delicate orchestration between technological design and its protective armoury.

- **Blueprinting Security Infrastructure:** The essence of any secure technological infrastructure starts with its blueprint. Drafting comprehensive policies and procedures for aspects like network security, device management, and system administration lays the groundwork for a fortified infrastructure. It's akin to building a structure with security ingrained in its very architecture.
- **Fortifying Network Structures:** With the ever-increasing complexity of networks, a strategic approach to their security becomes crucial. Network segmentation, for instance, helps in isolating different parts of the institution's network, thereby minimizing potential breach impact. Such an approach ensures that if one segment faces issues, the entire network doesn't come crashing down.
- **Strengthening Access and Oversight:** While the institution's technological assets are vital, it's equally essential to regulate who has access to them. Rigorous access control mechanisms act as gatekeepers, ensuring only authorized personnel can interact with critical systems. Coupled with real-time monitoring, institutions can maintain a vigilant eye, ensuring anomalies are detected and dealt with swiftly.

To encapsulate, this segment emphasizes the harmony between technological structuring and its protective layers. By fusing meticulous technology architecture with state-of-the-art security mechanisms, academic institutions can cultivate a digital environment that's as robust as it is secure.

Business Architecture		**Data Architecture**	
Business Strategy Alignment	**Plan**	Data Classification	**Plan**
Business Process Analysis	**Monitor**	Data Access Control	**Identify**
Cybersecurity Governance	**Review**	Encryption	**Detect**
User Awareness Training		Data Retention & Disposal	**Protect**
Compliance & Regulations		Backup & Recovery	**Recover**
Incident Response Plan		Data Leak Prevention	**Review**
Vendor & Third-party Management		Database Security	
Business Continuity & Disaster Recovery		Data Integrity Checks	
Resource Allocation		Audit Trails	
Stakeholder Communication Plan		Data Privacy Regulations	
Information Architecture		**Technology Architecture**	
Secure Development Lifecycle	**Plan**	Network Segmentation	**Plan**
Application Vulnerability Assessment	**Identify**	Firewall Implementation	**Identify**
Patch Management	**Detect**	Intrusion Detection &Prevention(IDPS)	**Detect**
Access Controls	**Protect**	Endpoint Protection	**Protect**
Application Performance Monitoring	**Recover**	Physical Security	**Recover**
API Security	**Review**	Network Monitoring & Logging	**Review**
Code Review		Access Controls	
End-point Security		Cloud Security	
Malware & Antivirus Scanning		Virtualization Security	
Application Training		Wireless Network Security	
INTEGRATE WITH PROPRIETORY FRAMEWORKS			

Figure 2: Proposed EA-based Cybersecurity Process

The process being proposed involves using all EA domains, as depicted in figure 2 above, to guide the design and implementation of strategies and security controls within higher education institutions in South Africa. Incorporating these diverse components within an EA-based cybersecurity blueprint empowers academic institutions to craft a holistic, robust, and agile security process. Such an architecture not only stands firm against evolving cyber threats but also seamlessly dovetails with the institution's overarching objectives and the ever-shifting educational environment. By fostering this symbiotic relationship between cybersecurity and

institutional aspirations, establishments can fortify their digital frontiers while staying in tune with their evolving operational paradigms. The proposed process can also be used with any other framework of choice.

6. The Role of Relationships Across all EA Components

EA provides a structured and comprehensive approach to aligning cybersecurity measures with an organisation's holistic objectives, infrastructure, and operations (Trad, 2023). Relationships between different components of EA are critical in identifying interconnections and vulnerabilities that cyber attackers can exploit. The following are the roles of relationships in EA components:

6.1 Identifying Interconnections and Vulnerabilities

EA serves as a keen eye for uncovering the intricate interplay among diverse facets of an organization's operational machinery, affording stakeholders a profound understanding of the collaborative efforts that steer the organization toward its goals, as elucidated by (Trad, 2023). This revelation of interconnections not only unveils the inner workings but also unveils potential points of vulnerability that cunning cyber criminals might seek to exploit.

To illustrate, should a vulnerability rear its head within one facet of an organization's operations, its repercussions can reverberate through the entire ecosystem, giving rise to a cascade effect that could culminate in a far-reaching security breach. Such is the power of comprehending these interconnections, as revealed through the lens of EA.

6.2 Analysis of Potential Attack Surfaces

EA lends its capability in delving into the realm of prospective attack surfaces, adeptly identifying the pivotal assets, integration between assets, and data that demand a vigilant shield of protection, as elucidated by (Masi et al., 2023). Through the intricate cartography of an organization's intricate processes and the graceful meandering of information, EA unfolds its capabilities in uncovering potential junctures of vulnerability and gracefully choreographing contingency plans. This comprehensive analysis emerges as a strategic compass, empowering organizations to judiciously prioritize their cybersecurity endeavours and efficiently allocate precious resources.

6.3 The Importance of Relationship Analysis in Cybersecurity

Relationship analysis is critical in cybersecurity risk management as it helps identify potential vulnerabilities and attack surfaces (Walter et al., 2023). Understanding the relationships between different components of an organisation's operations makes it possible to develop a comprehensive and integrated cybersecurity strategy that aligns with the organisation's goals and objectives. This approach ensures that cybersecurity measures are not just reactive but integrated into institutional operations' foundational structure. The diagram in Figure 3 below gives a view of the comprehensive nature of elements covered by EA.

When we decode the symbiotic ties between diverse operational components, it empowers organizations to draft a cybersecurity blueprint that is not merely reactionary. Instead, it becomes an integral part of the organization's foundational matrix, harmonizing with its overarching aspirations. The accompanying figure 3 (Archimate, 2000), drawing inspiration from the comprehensive tenets laid out by EA further explains this intertwined nature.

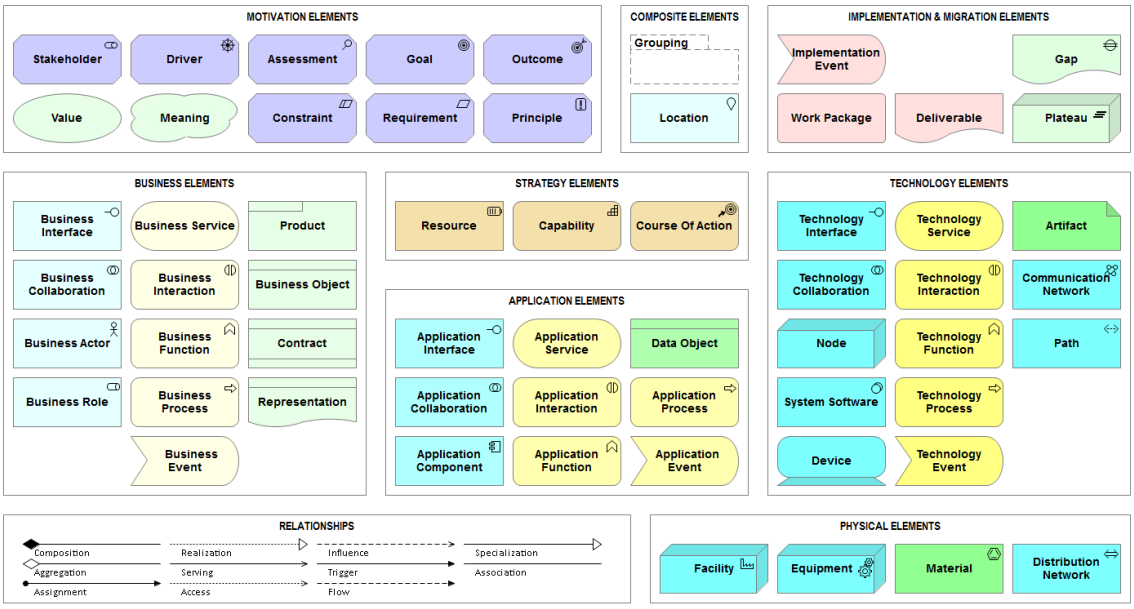


Figure 3: EA components – (Archimate, 2000),

The relationship within all EA elements can help us spot weak points that cunning attackers might want to exploit. By closely looking at these relationship and possible threats, organizations can better plan their online safety measures. This means they will not just respond to dangers, but they will also strengthen their online spaces in a way that fits their main goals.

6.4 Implications for Higher Education Institutions in Africa

We believe the EA-driven cybersecurity approach brings profound implications for universities, especially African higher education institutions, sparking a transformative shift. By aligning cybersecurity strategies with the intricate tapestry of academic goals, infrastructure, and operations, this approach promises heightened resilience against looming cyber threats. Moreover, it serves as the linchpin for seamless business continuity while harmonizing cybersecurity efforts with the overarching institutional vision.

Yet, the EA approach extends its reach further, offering a potent response to unique challenges facing African institutions—constraints in finances, a shortage of cybersecurity expertise, and pervasive user awareness gaps. Furthermore, it aids in navigating the complex legal landscape surrounding cybersecurity in many African nations, ensuring compliance with comprehensive legal frameworks in this domain.

7. Conclusion

The presented EA-driven cybersecurity approach offers a holistic and integrated methodology for harmonizing cybersecurity strategies with the comprehensive objectives, infrastructure, and operational facets specific to African higher education institutions. This approach carries notable implications for these institutions as they grapple with distinctive challenges in the implementation of effective cybersecurity measures. Prospective avenues for future research within this academic domain may encompass the development of a comprehensive legal framework, the augmentation of cybersecurity expertise, the pursuit of additional case studies, and the exploration of emerging technological paradigms.

References

Ahmed, M., Panda, S., Xenakis, C., & Panaousis, E. (2022). MITRE ATT&CK-driven cyber risk assessment. Proceedings of the 17th International Conference on Availability, Reliability and Security,

Archimate. (2000). Archimate EA Digram [Web]. https://www.archimatetool.com/resources/. https://www.archimatetool.com/resources/

Bouke, M. A., Abdullah, A., ALshatebi, S., Atigh, H. E., & Cengiz, K. (2023). African Union Convention on Cyber Security and Personal Data Protection: Challenges and Future Directions. *arXiv preprint arXiv:2307.01966*.

Burkett, J. S. (2012). Business security architecture: weaving information security into your organization's enterprise architecture through SABSA®. *Information Security Journal: A Global Perspective, 21*(1), 47-54.

Choudhuri, B., Srivastava, P. R., Mangla, S. K., & Kazancoglu, Y. (2023). Enterprise architecture as a responsible data driven urban digitization framework: enabling circular cities in India. *Annals of Operations Research*, 1-29.

De Haes, S., Van Grembergen, W., & Debreceny, R. S. (2013). COBIT 5 and enterprise governance of information technology: Building blocks and research opportunities. *Journal of Information Systems, 27*(1), 307-324.

Dlamini, I., Taute, B., & Radebe, J. (2011). *Framework for an African policy towards creating cyber security awareness*.

Eltahir, M., & Ahmed, O. (2023). Cybersecurity Awareness in African Higher Education Institutions: A Case Study of Sudan. *Inf. Sci. Lett, 12*.

Gloria Appiah, J. A.-A., Yu-Lun Liu. (2020). Organizational Architecture, Resilience, and Cyberattacks. *IEEE Transactions on Engineering Management, 69*(5, October 2022), 2218 - 2233.

Groš, S. (2021). A critical view on CIS controls. 2021 16th International Conference on Telecommunications (ConTEL),

Humphreys, E. (2016). *Implementing the ISO/IEC 27001: 2013 ISMS Standard*. Artech house.

IGNAT, I. (2022). *Factor Analysis of Information Risk (FAIR™) when assessing the Information Security* Universitatea Tehnică a Moldovei].

Kayode-Ajala, O. (2023). Establishing Cyber Resilience in Developing Countries: An Exploratory Investigation into Institutional, Legal, Financial, and Social Challenges. *International Journal of Sustainable Infrastructure for Cities and Societies, 8*(9), 1-10.

Kotusev, S., Kurnia, S., & Dilnutt, R. (2023). Enterprise architecture artifacts as boundary objects: An empirical analysis. *Information and Software Technology, 155*, 107108.

Lallie, H. S., Thompson, A., Titis, E., & Stephens, P. (2023). Understanding Cyber Threats Against the Universities, Colleges, and Schools. *arXiv preprint arXiv:2307.07755*.

Masi, M., Sellitto, G. P., Aranha, H., & Pavleska, T. (2023). Securing critical infrastructures with a cybersecurity digital twin. *Software and Systems Modeling, 22*(2), 689-707.

Maulana, Y. M., Azmi, Z. R. M., & Arshah, R. A. (2023). Modeling of Strategic Alignment to Modify TOGAF Architecture Development Method Based on Business Strategy Model. *International Journal on Advanced Science, Engineering & Information Technology, 13*(1).

Maulana, Y. M., Azmi, Z. R. M., & Phon, D. N. E. (2023). Business-IT Alignment through Enterprise Architecture in a Strategic Alignment Dimension: A Review. *Register: Jurnal Ilmiah Teknologi Sistem Informasi, 9*(1), 55-67.

Morse, E. A., & Raval, V. (2008). PCI DSS: Payment card industry data security standards in context. *Computer Law & Security Review, 24*(6), 540-554.

Nasir, N. N. I., Radzuan, S. N., Azhami, B. A., & Hamidon, H. (2023). Cyber Security in Higher Education: Problem and Solution. Proceedings of 1st Glocal Symposium on Information and Social Sciences (GSISS) 2023,

Pereira, C. M., & Sousa, P. (2005). Enterprise architecture: business and IT alignment. Proceedings of the 2005 ACM symposium on Applied computing,

Rana, M., & Sharma, D. (2023). Understanding Cyber-Attacks and their Impact on Global Financial Landscape. 2023 International Conference on Circuit Power and Computing Technologies (ICCPCT),

Rozas, I. S., Khalid, K., Yalina, N., Wahyudi, N., & Rolliawati, D. (2020). Digital Enterprise Architecture for Green SPBE in Indonesia. *CCIT (Creative Communication and Innovative Technology) Journal, 15*(1), 26-24.

Sharkov, G. (2020). Assessing the maturity of national cybersecurity and resilience. *Connections: The Quarterly Journal, 19*(4), 5-24.

Taherdoost, H. (2022). Understanding cybersecurity frameworks and information security standards—a review and comprehensive overview. *Electronics, 11*(14), 2181.

Trad, A. (2023). Integrating a Holistic Enterprise Architecture Pattern: A Proof of Concept. In *Handbook of Research on Digitalization Solutions for Social and Economic Needs* (pp. 1-39). IGI Global.

Walter, M., Heinrich, R., & Reussner, R. (2023). Architecture-based attack path analysis for identifying potential security incidents. European Conference on Software Architecture,

White, G. B., & Sjelin, N. (2022). The NIST cybersecurity framework. In *Research Anthology on Business Aspects of Cybersecurity* (pp. 39-55). IGI Global.

A Study into Privacy and Legal Issues in Cloud Computing: The Mozambican Context

Ambrósio Patricio Vumo
Universidade Pedagogica, Maputo, Mozambique
Technische Universität Dresden, Department of Computer Science, Germany

ambrosio_patricio.vumo@mailbox.tu-dresden.de

Abstract: During the last 10 years, Cloud computing has become an evolving technology providing several benefits such as cost reduction and high flexibility. However, one of the main challenges related to cloud computing is related to data security and privacy. Despite this, worldwide many countries, specially developed countries have adopted cloud computing technology. In Africa, specifically in the Southern African Development Community (SADC), Mozambique is one the countries who has recently adopted cloud computing technology. However, on the contrary to countries such as Mauritius and South Africa, Mozambique still does not have a national strategy for cloud computing in place, including security and privacy issues. International organizations such as ENISA and NIST as well as ITU have published frameworks related to cloud computing adoption covering data security and privacy issues. Therefore, in this paper we first analyze the cloud computing frameworks published by these international organizations. In addition, the paper also analyses the adoption of cloud computing in developed and developing countries, such as USA, UK, Germany, Mauritius and South Africa. From these analyses, the paper presents some recommendations for Mozambique to adopt best practices and follow international frameworks related to cloud computing including data security and privacy.

Keywords: Cloud computing, Data protection, PRIVACY, LAws, Mozambique

1. Introduction

Cloud computing and the emergence of a cloud economy based upon it is becoming increasingly important considerations for governments and enterprises (UN, 2013). A recent study has described cloud technology as one of the most significant disruptive technologies which will develop over the next two decades, with major implications for markets, economies and societies. It predicts that, by 2025, most information technology - IT and web applications and services could be cloud delivered or cloud enabled, and that most enterprises might be using cloud facilities and services. In simple terms, cloud computing enables users, through the Internet or another digital network, to access a scalable and elastic pool of applications, data storage and computing resources (Nelson Fonseca, 2015). Despite of the numerous benefits that cloud computing promises, its adoptions face numerous challenges such as (ITU, 2012): i) data security, the trust, reliability and dependency in moving data and applications to a remote third party; ii) availability of connectivity and sufficient bandwidth; iii) legal and regulatory uncertainty; and iv) in additional in developing countries there are varying challenges in their socio-economic and political development leading to low investments in advanced technologies for government, academic, industry and private. Many African countries have seen development and integration with the global economy limited due to lack of high performing technological infrastructures. Cloud computing has emerged as a new computing paradigm that will revolutionize the way we buy and use computing equipment and services in many countries of the African continent (UN, 2013). While it is always difficult, even risky, to predict the development of IT so far ahead, policy makers and business leaders should start examining the opportunities and potential risks of this growing phenomenon. The Internet infrastructure and connectivity today are providing a ubiquitous access to shared applications, software and hardware across the globe allowing users to quickly and easily get information whenever and wherever they might be. Cloud computing is taking this ubiquitous access to another level be providing users with access not just to information available on the Internet but also to their business applications and network wherever there is Internet connectivity using any compatible device such as a personal computer, laptop, tablet or smartphone. Users can have access to programs, storage, applications, processing and software development environments that are provided through cloud computing. The best of our knowledge there are no articles or research works on data security and privacy protection in cloud computing environment in Mozambique.

2. Theoretical Background

2.1 Cloud Computing Definition and Models

The National Institute of Standards and Technologies (NIST) of the United States of America, the International Organization for Standardization (ISO), International Electrotechnical Commission (IEC) together with the

International Telecommunication Union (ITU) Telecommunication Standardization Sector (ITU-T) are widely used to define Cloud computing where both definitions establish the fundamental nature of the concept such as "self-services" and "on-demand" through "access network":

- **NIST (**NIST, 2019), defines cloud computing as a model for enabling ubiquitous, convenient, on demand network access to a shared pool of configurable computing resource such as networks, servers, storage, applications and services that can be rapidly provisioned and released with minimal management effort or service provider interaction;
- **ISO/IEC together with ITU-T** (ITU, 2014), define cloud computing as a paradigm for enabling network access to a scalable and elastic pool of shareable physical or virtual resources with on demand service provisioning and administration.

According to (NIST, 2019) (ITU, 2014), there are three classic or three well-known cloud service models in the cloud computing paradigm, namely: i) Software as a Service -SaaS; ii) Platform as a Service - PaaS; and iii) Infrastructure as a Service - SaaS and four deployment models, namely: i) private cloud; ii) public cloud; iii) Community cloud; and iv) hybrid cloud.

2.2 Contributions and Opportunities of Cloud Computing

According to Varnika (Varnika, 2018), the on-going adoption of cloud computing services and potential for the future will have significant impacts on wide-ranging aspects of developing economies. Varnika stated that the cloud computing in the developing world is in its infant stage and is an attractive proposition to organizations due to the benefits that come with adopting it, they include (Varnika, 2018):

- **Enterprises Applications**: Utilizing the cloud significantly reduces the barriers to entry for entrepreneurs in the developing world by providing customized and scalable computing, storage, and development solutions without the need for significant capital outlays. Developing businesses can purchase only what they need at the time while being able to accurately forecast future costs in the face of growth. Similarly, developers and software engineers in these regions benefit from access to the same computing power availability as those they are competing with in the West, allowing them to produce applications and services that are better tailored to their local and regional market due to their increased understanding of the cultural and economic needs of the area.
- **Health Applications:** Cloud technologies are able to match patients with medical providers in ways that were previously impossible at the turn of the millennium. In areas that are often underserved by qualified medical facilities and providers, the future potential of the cloud to jump start the implementation of telemedicine is enormous. Effective telemedicine would allow providers to access remote patients, track disease outbreaks, and reduce the burden on physical hospital resources by prioritizing care.
- **Government Applications:** The most striking example of the benefits of cloud technology is the reduction in infrastructure required, including electrical, computing, storage requirements. Similar to applications in healthcare record management, the utilization of cloud computing and storage could greatly expand the quality, reliability, and affordability of government services. This would also greatly improve the redundancy of record keeping in areas prone to conflict, reducing the time required to return to normalcy following fighting.

2.3 Defining Security, Privacy, Trust and Personal Data

Nayak (Nayak, 2012) within his research entitled understanding the security, privacy and trust challenges of cloud computing has defined security, privacy and trust as following:

- Privacy concerns the expression of or adherence to various legal and non legal norms regarding the right to private life. In the European context this is often understood as compliance with European data protection regulations. Although it would be highly complex to map cloud issues onto the full panoply of privacy and personal data protection regulatory architectures, the globally accepted privacy principles give a useful frame: consent, purpose restriction, legitimacy, transparency, data security and data subject participation (Nayak, 2012).

- Trust revolves around assurance and confidence that people, data, entities, information or processes will function or behave in expected ways. Trust may be human to human, machine to machine, human to machine or machine to human. At a deeper level, trust might be regarded as a consequence of progress towards security or privacy objectives (Nayak, 2012).
- Security concerns the confidentiality, availability and integrity of data or information. In accordance with this author, security may also include authentication and non-repudiation (Nayak, 2012).
- Personal Data is defined as information of any kind referred to identified and/or identifiable individuals and/or entities. It is a very broad definition which encompasses almost any sort of information linked to a subject. In turn, data treatment is also widely defined as the systematic operations or procedures, by electronic means or others, allowing storage, conservations, modifications, relationship, evaluation, lock, destruction and in general data processing of personal data, as well as assignment to third parties through communications, interconnections or transfer (Lisandro Frene, 2019).

2.4 Challenges of Data protection to address new technologies

According to (UN, 2013), Data protection is a dynamic field that is constantly challenged and influenced by advances in technology and innovation in business practices. The relationship between data protection and ICT developments changes all the time and can be demonstrated by three recent developments: i) Cloud computing; ii) the Internet of Things; and iii) Big Data analytics. Each of these technologies presents new challenges to data protection, particularly in the areas of the definition of personal data and the management of cross-border data transfers.

3. Privacy and Legal Issues in Cloud Computing

3.1 Privacy Issues in Cloud Computing

According to (Benameur, 2011) (Dlodlo, 2011), privacy is a fundamental right enshrined in the UN Universal Declaration of Human Rights. There are various forms of privacy, including the right to left alone and control of information about ourselves. There are different types of information that need to be protected. These include any information that can be used to identify or locate an individual such as Name, address, credit card number and IP address. Sensitive information such as personal financial information and job performance information is considered private. Behavioral information such as viewing habits for digital content, users recently visited websites or product usage history need to be protected as well. According to Siani and Azzedine (Benameur, 2011) in the commercial, consumer context, privacy entails the protection and appropriate use of the personal information of customers, and the meeting of expectations of customers about its use. For organizations, privacy entails the application of laws, policies, standards and processes by which Personally Identifiable Information (PII) of individuals is managed. According to (SG, 2019) data privacy is one of the most important components which individuals/business organizations take into consideration before hiring the services of any third-party cloud computing service provider. Consumers and businesses are willing to use online computing only if they trust that their data will remain private and secure. Privacy and confidentiality are regarded as the major problems in cloud computing. Individuals to a large extent do not have any problem in sharing their information with the cloud service provider, but when it comes to business organizations, they are very anxious about sharing proprietary data. There are certain categories of businesses that are by specific laws prohibited from sharing certain data. For instance, for a healthcare company in the United States, the Health Insurance Portability and Accountability Act (HIPAA) lays down certain restrictions in relation to the sharing of medical records of individuals. In cloud computing environment, services can be aggregated and changed dynamically by customers and service providers can change the provisioning of services anytime. Nomusa (Dlodlo, 2011) has stated that violation of privacy occurs as a result of a number of cloud computing dynamics. In the cloud computing the infrastructure is shared between organizations and is off-premise. Therefore, there are threats associated with data being stored remotely and because of virtualization. Virtualization is a method of running multiple independent virtual systems on a less physical resource making on computer act as many, and sharing the resources of hosts across multiple environments. Sensitive data may move around within an organization and across organizational boundaries. Thus, Legal compliance and adequate protection has be maintained therefore.

3.2 Legislative and Regulatory Issues in Cloud Computing

Nomusa (Dlodlo, 2011) stated that in cloud computing there are a number of issues that need to be regulated. Potential physical location of data centers could be anywhere, with geography-blind distribution of applications and data. As a practical commercial matter, national regulations should be able to influence the actual deployment of cloud computing services in countries around the globe. SG (SG, 2019) also stated that Legislative and regulations is another aspect of cloud computing which is largely uncertain as cloud computing is an unregulated field with a mixture of legislation and regulations present in limited countries. The problem arises when the jurisdiction where the customer resides requires a certain standard for the data which may not be required by the jurisdiction where the data centre is located. For instance, the customer's country might have certain data privacy, data retention and data access enforcement laws, but these might not be the same as the laws of the service provider's country, and the laws might conflict with each other. For instance (Dlodlo, 2011), in the US and Europe the regulations require some cloud computing offerings to allow users to stipulate the country in which their data will be stored. Non-US firms whose servers are located in the US can have their information accessed by the US government under the US Patriot Act and Homeland Security Act. This impact on information privacy policy.

3.3 Data Location, Law and Jurisdiction Issues in Cloud Computing

According to (sg, 2019), in cloud computing environment the location of data is usually uncertain. Customers at times are unaware as to the region where the cloud infrastructure is located. The data stored or processed by any individual or business organization travels over the internet network to various data centers located in multiple locations of the world. This gives rise to various legal issues as the cloud infrastructure is located elsewhere, not in the country where the customers reside, and different countries have different laws pertaining to data storage/transfer. Thus, it is the contract of service that determines the applicable law and the courts who should decide matters in the case of disputes. With cloud computing, the applicable laws governing your data could include the laws where your organization is headquartered, where your cloud provider is headquartered, where your cloud provider's data centers are located, where the subjects of the data reside, and potentially the laws of the countries that your data pass through on their way to, from and among the cloud provider's data centers. For these reasons it is essential for a cloud computing contract to identify the geographic region within which the data centers hosting your data, and potentially the headquarters of the cloud provider, may be located, and to specify the cloud provider's obligations to keep your data in those regions. Otherwise, the overlaps and potential conflicts between the possible governing laws could make legal and data access compliance impossible. According to [UN], Data protection regulation is high on the political agenda as evidenced by a number of current developments:

- The United Nations in 2015 appointed a Special Rapporteur on the right to privacy.
- The African Union in 2014 has developed the African Union Convention on Cyber security and Personal Data Protection.
- A new General Data Protection Regulation has replaced the European Directive on Data Protection, which has been a prominent source of regulation for twenty years.
- Data protection has been included in several international trade agreements.
- Data protection regulation has been considered in several high-profile court cases in relation to national surveillance.
- Numerous countries are drafting new data protection laws or are reviewing existing ones.
- The European Union and the United Stated have re-negotiated a long-standing cross-border data protection agreement.
- Several global and regional organizations have issued (or are developing) multiparty agreements and /or guidelines on data protection.

4. Privacy and Legal Issues in Mozambique

According to Dla Piper (2024), in Mozambique there is no specific legislation on data protection or privacy. Vumo et al (2019) stated that the Constitution of the Republic of Mozambique outlines some aspects related to protection of personal data. For instance, pursuant to Article 41 of the constitution of republic of Mozambique

states that all individuals are entitled to intimacy of their private life (Joao Luis Traca, 2016). Further, Article 71 of the Constitution grants all individuals the right to privacy, prohibiting the use of electronic means for recording and processing individually identifiable data in respect of political, philosophical or ideological beliefs, of religious faith, party or trade union affiliation or private lives (Joao Luis Traca, 2016). Moreover, Data Protection is referred on the Electronic Transaction Law approved in 2017. But it provides only some degree of personal data protection and privacy. Still today, Mozambique does not have a full regime for data protection and privacy (UN, 2013). However, there are other sources of law that impose some privacy obligations such as: i) The Civil Code (Decree-Law no. 47344, of November 25, 1966, in force in Mozambique through Edict no. 22869, dated September 4, 1967); ii) The Penal Code (Law no. 24/2019, of December 24, as amended by Law no. 17/2020 of 23 December); iii) The Labour Law (Law no. 23/2007, of August 1) and the new Labour Law (Law no. 13/2023, of 25 August) which has entered into force on 22 February 2023; iv) The Regulations on Registration and Licensing of Intermediary Electronic Service Providers and Operators of Digital Platforms (Decree no. 59/2023, of 27 October) Dla Pipel (2024).

Despite the Constitution of the Republic of Mozambique outlines aspects related to data protection and privacy, Mozambique is among the least committed countries in cybersecurity within five designated areas, namely (ITU, ITU, 2019): i) legal; ii) technical; iii) organizational; iv) capacity building; and v) cooperation. In this context, Mozambique could be viewed as a permissive environment for cyber criminals due to a lack of security capabilities, absence of specific legislation on data protection and privacy and other cyber related legislations. Currently several governmental and non-governmental organizations have been discussing data protection and cyber security issues but there is a lack of coordination between these organizations and concrete actions are missing. Thus, due to the dynamic nature of cyberspace, there is a need for these actions to be unified under national level with an integrated vision and a set of sustained and coordinated strategies for their implementation. Despite this cloud computing is reality in Mozambique. However, the absence of relevant laws and regulations to regulate this environment for data storage, conservations, modifications, lock-in, destruction and in general data processing of personal data, as well as assignment to third parties through communications, interconnections or transfer could pose risks and lack of trust to cloud consumers.

5. Addressing Privacy and Legal Issues in Mozambique

5.1 National Considerations

As previously presented issues related to privacy, trust and legal issues in cloud computing and based in theoretical evidence, we present some initial steps that help address these concerns. We proposed that, the government of Mozambique should:

- **Develop Laws:** Jacques Rousseau (2012) explained in the social contract, or principles of political right, the rules that the members of a society create to balance the individual rights to self-determination against the needs of the society as a whole are called laws. Laws are rules mandate or prohibit certain behavior; they are drawn from ethics, which define socially acceptable behaviors. In Cloud computing environment, laws are needed to governing the issues related to data protection and privacy. This can be addressed mainly by the geographical location of the stakeholders involved, and the rights and obligations of each stakeholder are determined by applicable regulations. In this context, it is very important to develop specific Personal Data Protection Law.
- **Include the core principles of Data protection:** According to (UN, 2013), recommend for countries that do not yet have laws in place, or countries that are updating or reforming their laws, should look to include the core principles in their new/amended legislation as following: i) openness, organization must be open about personal data practices; ii) collection limitation, collection of personal data must be limited, lawful and fair; iii) purpose specification, purpose of collection and disclosure must be specified; iv) use limitation, use of data must be limited to specific purposes; v) security, personal data must be subject to appropriate safeguards; vi) data quality, personal data must be relevant, accurate and up-to-date; vii) access and correction, people must be able to access and correct their personal data; and viii) Accountability, data controllers must comply with the data protection principles. Accountability in the cloud is a very important concept that needs to be supported from both a legal and technical viewpoint. The way to achieve accountability in such a dynamic and worldwide infrastructure is to have a strong emphasis on auditing. Audit should be able to keep track of where the data has been outsourced, who processed it and for what purpose.

- **Establish Data Protection Authority:** The data Protection Authority is an independent supervisory authority ensuring that privacy is protected when personal data are processed. Ensuring that a cloud deployment meets the requirements imposed by the applicable laws, regulations and contractual obligations, the National Data Protection Authority must take responsibility for ensuring compliance with the data protection principles as recommended by AU (2020).

5.2 International Considerations

According to (UN, 2013), Cloud computing services do not present unique issues in data protection, but they do add to the complexity of existing issues, especially in relation to cross-border data transfers. To date, few jurisdictions have attempted to draft regulations expressly designed to regulate the provision of cloud computing services. This probably reflects both the broad range of services that fall within the concept of cloud computing, as well as the flexibility of scope within existing regulatory concepts. In this context, we proposed that, the government of Mozambique should:

- **Regulate international data transfers:** States and businesses often have a legitimate need to share data across national borders for economic and logistical purposes. This must be balanced with the protection of privacy and data security of the data of their citizens. It is important for frameworks to consider how to institute accountability mechanisms to ensure that those processing and using the data are accountable (UNDP, 2023). Thus, Data protection law must expressly regulate international data transfers, allowing it if the requirements set forth by law for such purpose is met. Requirements vary, essentially according to the purpose of the data transfer; and to the country where data is exported. GA Solanki (Solanki, 2012) also stated that the traditional rules of private international law, the jurisdiction of a nation only extend to individuals who are within the country or to the transactions and events that occur within the natural borders of the nation. However, this traditional rule pertaining to jurisdiction has become less effective with the advancement of commerce and technology. In cases of cloud transactions, it may happen that a company which is resident of one country may stored data on a cloud which is located in altogether different country and such cloud may belong to a vendor who is located in a third country. In such cases there are ample chances that the laws of third jurisdictions are applicable. Therefore, various factors would need to be considered while determining an appropriate jurisdiction along with the harmonization of domestic laws of each applicable country to avoid conflict of laws. We also propose the Government of Mozambique to keep cloud data compliant with global regulations (GDPR, CCPA, HIPAA, PCI DSS and FISMA) and internal policies as stated by (Sentra, 2024).

6. Conclusion

Despite the fact that cloud computing provides several services to individuals and organizations alike, there are also many problems that need to be addressed and one of them is related to the way how information is managed. As discussed in Section 4 data security and privacy protection is the main concern in cloud computing environment. This is due to the characteristics of the cloud infrastructure because cloud computing services are often owned by a third party. Moreover, without knowledge of the physical location of the servers by users, user's data in the cloud computing environment are easier to manipulated leading to the loss of control. In this paper, we have examined the challenges of data protection and privacy in cloud computing. The study was based from the following sources, UA, NIST, UN, ITU and UNDP, just to enumerate a few. The findings from this study show that despite some efforts, data protection and privacy, especial in cloud computing is still a challenging process in Mozambique. To date, there are no data protection authority and specific legislation on data protection or privacy. Therefore, the paper made some recommendations for the government of Mozambique in order to change the current situations as described on section 5. However, we believe that more efforts should be done by the Mozambican government to provide trusted, safe and protected cloud computing environment to its citizens and the society in general.

References

Ambrosio, Vumo and Josef, Spillner. (2019). A Data Security Framework for Cloud Computing Adoption: Mozambican Government Cloud Computing. *18th European Conference on Cyber Warfare and Security*, ACPI, pp. 720.

AU, (2020) "African Union Convention on Cyber Security and Personal Data Protection", [online], African Union, https://au.int/en/treaties/african-union-convention-cyber-security-and-personal-data-protection.

Siani, Pearson and Azzedine, Benameur. (2011). "Privacy, Security and Trust Issues Arising from Cloud Computing", *International Conference on Cloud Computing Technology and Science*, IEEE.

Dla Pipel, (2024) "Data Protection Laws of the World: Mozambique", [online], Dla Piper, https://www.dlapiperdataprotection.com/.

Dlodlo, N. (2011). "Legal, privacy, Security, Access and Regulatory Issues in Cloud Computing", *Interntional Conference on Information Management and Evaluation*.

ITU, (2014) "Information technology - Cloud computing - Overview and vocabulary", [online] https://www.itu.int/rec/T-REC-Y.3500201408-I

ITU. (2012). *Demystifying Relation in the Cloud: Opportunities and Challenges for Cloud Computing.*

ITU. (2024), "Global Cybersecurity Index 2020", [online], ITU, https://www.itu.int/epublications/publication/D-STR-GCI.01-2021-HTM-E.

Joao, Luis Traca and Lidia, Neves. (2016). "Data Protection in Mozambique: Inception Phase. *African Data Privacy Laws, Law, Governance and Technology"*, Spring International Publishing, pp. 363-367.

Lisandro, Frene and Juan, Cardinal (2019). *Data Protection and CyberSecurity.* Chambers: Global Practice Guide.

Michael, E. Whiteman and Herbert, J. Mattord. (2012). *Principles of Information Security.* Course Technology, Ed. 4, Boston.

Debabrada, Nayak. (2012), "Understanding the Security, Privacy and Trust Challenges of Cloud Computing". *Journal of Cyber Security and Mobility*, Vol.1, pp. 277-288.

Nelson, Fonseca and Raouf, Boutaba. (2015). "*Cloud Service, Networking, and Management*", New Jersy: John Wiley & Sons.

NIST. (2019). "*The NIST definition of Cloud Computing" [online]*, NIST, https://www.nist.gov/publications/nist-definition-cloud-computing.

Sentra. (2024) "Cloud Data Security: Challenges and Best Practices" [online], Sentra, https://www.sentra.io/cloud-data-security.

SG. (2019). *"*Data Privacy is one of the most Important Components*"* [online], SG, https://sg.inflibnet.ac.in/bitstream/10603/98806/14/14_chapter7.pdf.

Solanki, G. (2012). "Welcome to the Future of Computing: Cloud Computing and Legal Issues". *InternationalJournal of Scientific and Technology Research*, Vol. 1, pp. 30-33.

UN. (2013). *Information Economy Report 2013: The Economy and Developing Countries.*

UNDP. (2023), "Legal identity and data protection in modern societies", [online], UNDP, https://www.undp.org/blog/legal-identity-and-data-protection-modern-societies

Varnika, V. (2018). "Cloud Computing Advantages and Challenges for Developing Nations". *Interntional Journal of Scientific Research in Computer Science and Engineering*, ISE, Vol.6, pp. 51-55.

Masters Research Papers

Using Blockchain to Secure Digital Identity and Privacy Across Digital Sectors

Jaynill Gopal and Stacey Omeleze-Baror
Department of Computer Science, EBIT, University of Pretoria, South Africa

jgopal@tuks.co.za
Stacey.baror@cs.up.ac.za

Abstract: In the digital age digital data has been seen as the new currency for both companies and bad actors, leading to mismanagement of personal data by both entities. This mismanagement could lead to personal data being breached while this may prove to be beneficial to certain entities, this however is not the case for users and digital citizens. In recent years there has been an influx in data breaches and data mismanagement cases throughout various industries, including corporations such as Apple and Facebook, involving critical data relating to user's digital identities, personal data, and other identifiable information. This issue may be addressed by employing the use of blockchain technology, a technology that has been recognized as a secure system promoting security and privacy, we can prevent this mismanagement of data and data breaches from happening. While blockchain is a relatively new technology, there is potential to use it in the current age of the internet and digital identities by employing blockchain technology to secure one's digital identity. This research aims to propose a potential solution to the issue of data protection and data breaches by proposing and making use of a conceptual model which makes use of a multi-level blockchain system. The system isolates data from digital identities between various digital platforms to minimize the amount of data breaches that may occur – further minimizing the amount of personal data which may be breached. This system allows the user to have full control over who may access their private data, while preventing bad actors from accessing the data without the user's authorization. The multi-level blockchain splits the user's data according to industrial or digital sector in which their data is used, with a master blockchain acting as the connecting link to a person's digital identity. Making use of this multi-level blockchain allows for the user to control who has access to their data, while remaining anonymous and secure in a digital platform's database or digital storage.

Keywords: Digital identity, Blockchain, Data protection, Digital privacy, Privacy, Data breach

1. Introduction

The introduction of the internet to the public has led to many advantages with many people becoming accustomed to the versatility of the internet, with popular uses to the average user being social media, online banking, and shopping. While the internet makes the life of the average user easier, most share their information willingly online (contact details, and other identifiable information) and some of this information is repeated across those various platforms. The user believes that this information is kept safe and confidential by these online platforms, however there are threats to users' online identity which they are often unaware of, the greatest threat being data breaches (Neto et. al., 2021). A data breach is an event where an unauthorised entity gains access to user or client data stored in a digital platform's database, these events may be due to either internal or external entities (Mohammed & Tejay, 2023). A data breach could also be defined as data that is sold by those platforms. This threat could be dangerous to user data in the wrong hands, these wrong hands are not limited to just cyber attackers, but they also refer to those who use that information. Once this data is in the hands of malicious parties (or third parties), a user's digital identity is exposed and can be used to identify a user and target that user directly – either through marketing, phishing attacks, account high jacking, etc. While most of these may seem harmless, there could be a harmful impact to the user should their digital identity be made public to those unauthorised parties.

Since most users use the same login credentials to login to different digital platforms (Gao, et. al. 2020), this information can be considered as part of their digital identity and identifiable information. By protecting this information, we would be able to prevent targeted attacks on a user's digital identity. This paper introduces a method for protecting a user's digital identity using blockchain, a technology which is used to protect data and anonymise its users. Aiming to protect a user's digital identity by keeping the user anonymous between different digital platforms, while still being able to be identified by the platform as an anonymous digital identity.

2. Background Information

Blockchain technology, since its release in 2009, has been recognised for its security and reliability through its decentralised nature. Blockchain brings forth confidentiality, and privacy by using cryptography to secure the data which it contains. By replicating its data to nodes in a blockchain network, blockchain can achieve a

decentralised nature by bringing a reliability to the system – being resistant to disasters which centralised systems suffer from. In the case of blockchain if a node in the system goes offline, there is another available to serve the network (Dai, et. al., 2017). The architecture of blockchain consists of a chain of blocks that are stored in a node, a block stores data in the system, and a node is a computer system which holds a copy of the blockchain (Zhang, Xue, & Liu, 2019). Security in blockchain comes from the cryptography and immutability properties of the system, the decentralised nature of blockchain requires that all nodes need to have the same information. This is achieved by a consensus mechanism that blockchain employs, if one node differs from others then the node will be deemed invalid (Mohsin, et al., 2019). While blockchain is an attractive technology in terms of security, there are however drawbacks namely: storage, computability, and energy consumption. These drawbacks, depending on the use case, can be negligible if security is the main aim of the system that the blockchain is being used in. With privacy and security being one of the focus points of blockchain technology, anonymity is enforced – highlighting the privacy aspects of the blockchain, this is because blockchain uses hashed identifiers to store identities (Takemiya & Vanieiev, 2018).

A digital identity is an identifier which is used to identify a user over a computer network (Ante, Fischer, & Strehle, 2022). A user's digital identity contains information such as a user's login name, personal details, location information and other information depending on the platform for which the identity is hosted on (Saklikar & Saha, 2006). In recent years sharing personal information on digital platforms have become ever prevalent, this prevalence has brought a concern that users often do not consider the reprecussions of sharing their information online – they simply click on accept when digital platforms ask the user to review and accept the terms of using that specific platform without reviewing those terms (Chan & Virkki, 2014). This behaviour has been seen since the rise of social media and as the internet gained popularity in the past decade, this negligence for digital privacy has made it easier for digital platforms to process and analyse data without the user's knowledge. Processing of user data does not need to be done soley by the digital platform, as is common practice by digital platforms the user data collected is often sold to third parties who may or may not have malicous intentions (Saglam, Nurse, & Hodges, 2022).

The recent rise in reported data breaches over the past decade, with over 100 breaches being reported per year, has brought on a concern for information and data protection by netizens and government regulators alike. A famous breach being the 2014 iCloud breach, where private pictures of celebrities were leaked to the public because of this breach (BBC News, 2014). While this breach was not through an attack on Apple's iCloud infrastructure, but rather reported as targeted attack on the users' accounts by gaining unauthorised access using the victims' login credentials to gain access to those pictures. This unauthorised access was done due to the attacker acquiring the login ID of their intended victim, then by either brute forcing the victims' passwords, or using their security question answers to gain access to the accounts (Marwick, 2017). While preventing brute forcing passwords and answering security questions may be preventable or unavoidable, we can protect a user's login ID by protecting their digital identity.

Another example of a data breach is the Facebook Cambridge Analytica data scandal which came to light in 2018. The focus of this scandal was due to Facebook building up psychological profiles based on its users' interactions on and off the platform. These psychological profiles were then used against the users in the form of targeted advertisements (Hinds, Williams, & Joinson, 2020), while this is common practice in the digital realm, the data that Facebook obtained from those users', psychological profiles and other identifiable data were handed off to a third party – Cambridge Analytica (CA) – without the users' consent (Shipman & Marshall, 2020). In this case, CA was not bound by Facebook's terms and conditions agreement with Facebook's users and were seemingly free to do what they wanted with the data that Facebook had handed to CA. This form of data usage is common practice in big data (Hashimova, 2016) and is used in targeted digital advertising, this is due to advertisers being able to link a user's digital identity to an advertisement that is based on a user's digital history.

While there are authentication methods which do allow login using third-party platform's single-sign on (SSO) and OAuth which limit the need for resharing personal information (Sadqi, Belfaik, & Safi, 2020), there are some security and privacy concerns using SSO (Karie, et. al., 2020). The information shared to the platform using SSO is not in the user's control, using additional trackers attached to the user's session (Pham et. al., 2023).

3. Methodology

This paper introduces the Multi-layered blockchain (MLB) system model to act as a guide to protect users' digital identities on the internet. The methodology used in the proposed model would be a combination of

literature reviews pertaining to the research topic, the viability of blockchain usage in different industries in terms of security and data privacy to one's digital identity. The result being a model of a multi-tier blockchain system to facilitate the storage and separation of data to protect one's digital identity.

3.1 Multi-Layered Blockchain System Model

This model would demonstrate that a user would be the sole possessor of their information, with a sub-blockchain created for each digital platform that a person chooses to share their digital identity with. The blockchain effectively forms a pseudo hash-table using blockchain with each node being related to a digital platform that a user chooses to share their information with, and only have access to that node.

Each digital platform that a user has a profile on will have an internal blockchain of user profiles which contain a link to an identity in the Global blockchain (GB), the digital platform's blockchain would not contain personal information associated with the user's digital identity - only the user's account information associated with the digital platform would be stored. This would encapsulate and isolate important digital platform and user information, preventing access to that information outside of the digital platform. The global blockchain would be the basis of authenticating a user against a specific digital platform profile.

Figure 1 provides a visual representation of the MLB. As seen in GB, there are digital platform chains which connect to an identity. Each identity is stored in a separate blockchain and is logically separated from the digital platform that the user has an account with. The User Identity Blockchain stores personal and identifiable information which is often asked by websites, such as a user's name, contact information, and other identifiable information.

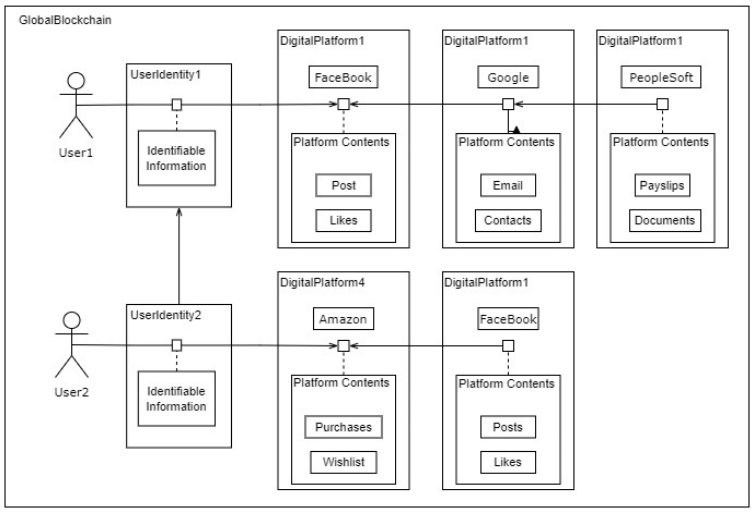


Figure 1: Visual Representation of the Multi-level Blockchain System

Figure 2 represents the internal blockchain that will be used by a digital platform. As seen each user profile has a link to a user identity within the global blockchain as depicted by the blocks *UserIdentity*. As seen in the *UserProfile* blocks, the information stored in the user profiles are only that of which is relevant to the digital platform. All non-platform specific user information is stored in the user identity.

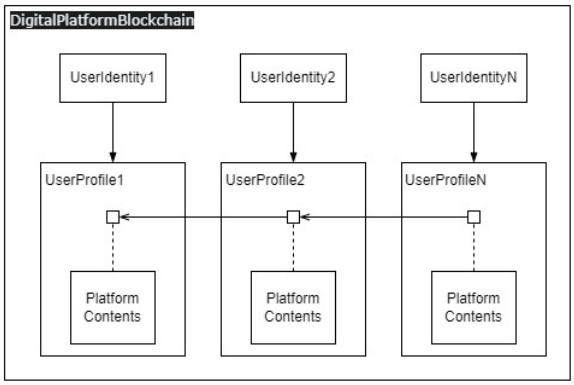


Figure 2: Visual Representation of the Digital Platform blockchain

3.2 Specification

The design of the MLB is a, as the name suggests, blockchain system with multiple blockchains. As mentioned in the previous section, the system consists of a Global blockchain, and an Internal blockchain. This approach provides two advantages – to prevent data duplication across digital platforms, and to ensure that should a data breach or leak on a digital platform occur the personal information of its users are not compromised in the breach. This is done by isolating the digital platform from its users, allowing the user a single place to store their digital information as well as to prevent attacks on digital platforms by attackers trying to exploit vulnerabilities on a digital platform to gain access to its user information.

3.2.1 Global blockchain

The Global blockchain is a public blockchain which acts as the registrar of all Digital Identities, where a single block in the system consisting of a single user's digital identity; since each user will have their own block, each block will contain a unique identity. The blockchain will use a hybrid technique depending on the operational scenario that would require the Global Blockchain. New blocks will require a Proof of Stake in order to be added to the chain, this will only occur when a user first creates their digital identity. New nodes will require Proof of Work, Proof of Authority, and Delegated Proof of Stake – this is to ensure that the node is of a trusted source. Transactions relating to specific blocks such as requests for information will require a Proof of Authority in order to validate the request and authenticate it accordingly.

A digital identity block would contain personal information about a user that is shared with a digital platform, information such as a user's contact information, geographic information, demographic information, financial information, or any other information which a user is likely to share to a digital platform. When a digital platform requests the user to access the information in the form of a claim, a request is sent to the user for the information. The user will need to review and authenticate the request before the information is sent to the digital platform in the form of a payload, which contains only the information that is requested by the digital platform. The full contents of an identity block will not be released to the digital platform, only the contents recognised by the claim that the digital platform has requested will be sent.

Requests for information are to be requested from a digital platform that has a valid claim granted by a trusted authority, unsigned or self-signed claims will not be accepted by the Global blockchain to be sent to the user for approval. Requests from an unregistered, or untrusted platform will be automatically denied by the Global Platform. Should a claim be successful and approved by the user, the requested information will be sent in an immutable format, with an appropriate time to live limitation. This is to prevent the digital platform from storing the information indefinitely as well as to prevent the information from being altered or distributed.

3.2.2 Internal blockchain

The Internal Blockchain is a private blockchain which is specific to the digital platform, this means that the blockchain may be implemented however the platform sees fit. The contents of the blockchain would consist of blocks containing information that is specific to the digital platform. Each block would contain a reference to the associated digital identity block in the Global blockchain, see Figure 2, as well as platform specific information. Since there is a reference to a user's digital identity in the Global blockchain, the internal blockchain does not need to – and should not – contain any personal information related to the user. If the platform requires the use of a user's personal information – such as the user's cell phone number – the platform should request that information via a claim.

A digital platform should register to apply for specific claims (depending on the category which the digital platform falls a part of) from a trusted authority, a digital platform can only apply for a claim which requests platform specific information – a social media platform may not request a user's credit card information. Should a digital platform request information via the claim and the request is approved, the platform will have a limited period to process that information before the authorisation would be revoked and the information will be discarded from the platform.

4. A use Case for the Multi-Layered Blockchain System

Suppose we have User1 (U1) who creates a profile on a digital platform (DP1) using the conventional methods of profile creation on digital platforms. U1 shares their name, email address, cell phone number and their social security number on the platform. The user uses the platform for a few months, soon after they discover that they are receiving many spam emails and phone calls than usual. The user is annoyed by this occurrence;

however, they think nothing of it. A few days later, they learn that DP1 has suffered a data breach where all account information of all its users including contact details and social security numbers, have been leaked a few days after U1 had created their account. After finding out about this incident, U1 decides to check if their social security number had been used after the leak and found that it is now associated with several credit accounts which U1 has no knowledge of. This mean that U1 would need to undo the damage caused by the entity or entities using U1's details, which could be a lengthy process. In this scenario the data breach had led to U1's identity being stolen, and their contact information being sold, as a result of the data breach on DP1.

The Multi-layered Blockchain system (MLB) aims to prevent this from occurring, suppose that the system is in use on a global scale. The global blockchain (GB) would contain all personal data needed for a user's digital identity, including their social security number, and contact details. Suppose we have User2 (U2) who created a profile on digital platform2 (DP2) using the MLB. The user would create their profile on DP2 by linking their digital identity to DP2. DP2 will only have access to the user's digital identity, and not the contents of the identity, Figure 3 illustrates how the processes used in the MLB.

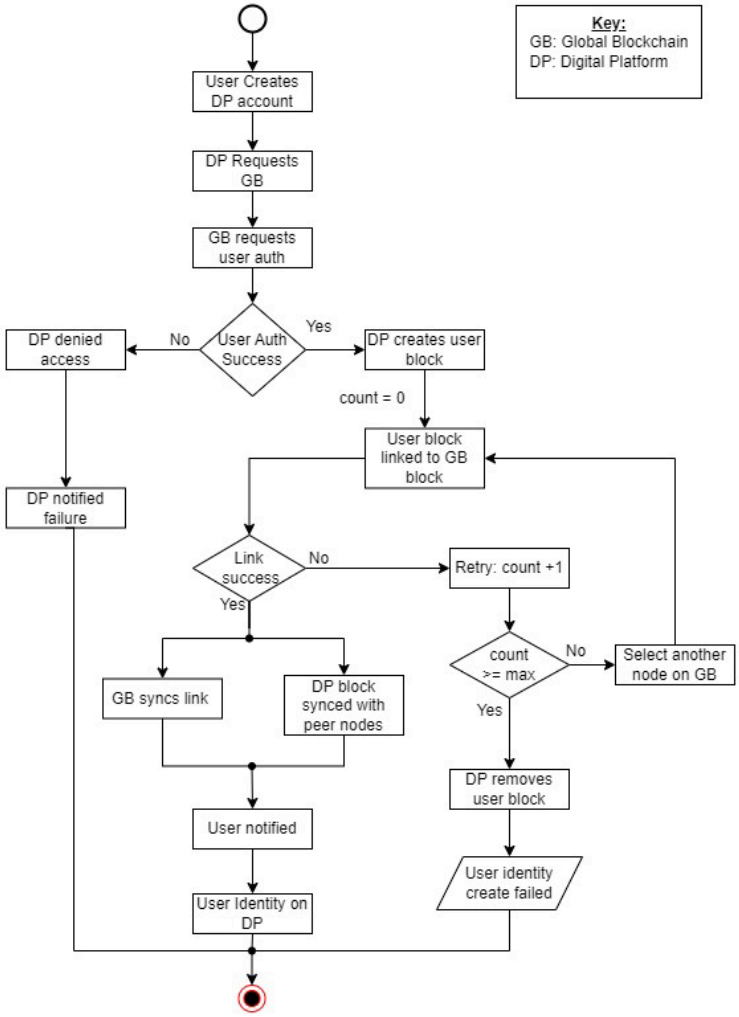


Figure 3: Process used in the Multi-layered blockchain

Once U2 chooses to link their digital identity to DP2, DP2 would send a request to the GB, which would request the user to approve this request to link their identity to DP2. Once this request has been approved by the user, U2's digital identity is subsequently linked to their profile on DP2. This link to DP2 will merely form as a link to the user, and DP2 will not have access to any of U2's personal information, Figure 4 represents the logical flow of the process involved should a digital platform request to access a user's information.

After some time, DP2 suffers from a similar data breach that DP1 had suffered from. In this incident however the data that had been breached was a list of profiles on DP2, with no accessible personal information to U2 – only the hash code containing the address of U2's digital identity on the GB was exposed. Sometime after the breach U2 has noticed a request to access information by another actor, which U2 has no recollection of

linking their digital identity to, U2 decides to deny this request for information as well as block any requests coming from this actor. This request will be denied by the Global blockchain, and any subsequent requests will result in the GB recognising them as an invalid request. In this incident, U2's information was protected by the MLB and as a result they did not suffer the same fate that U1 had suffered from.

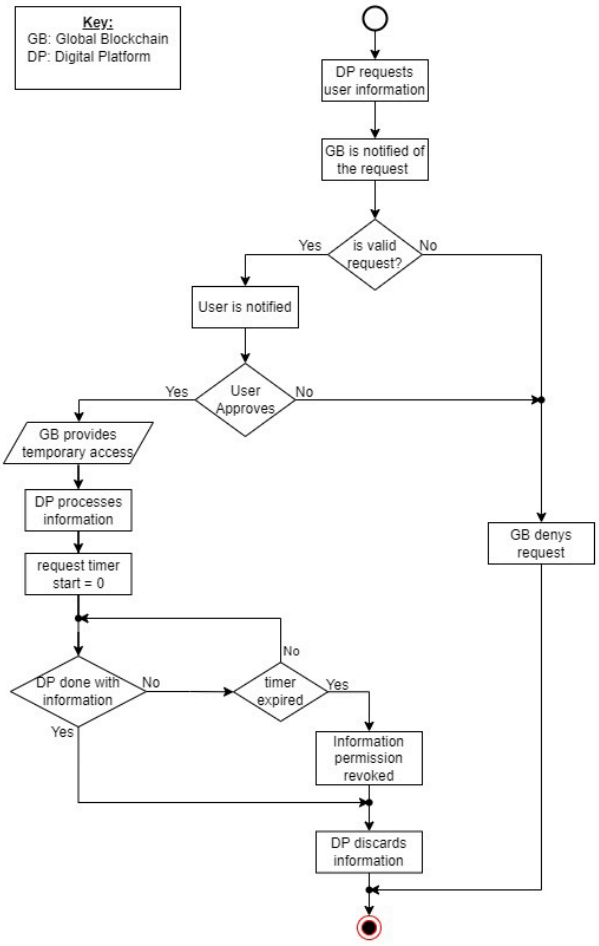


Figure 4: Request for information process using the Multi-layered Blockchain

5. Evaluation

The MLB model provides the user with a protected digital identity through storing their digital identity in a blockchain, while preventing data duplication between different digital platforms and risking their data from being compromised through data breaches on digital platforms. This allows the user's identity to be isolated from a digital platform, and any information stored about the user on is safely stored on the platform's internal blockchain. The evaluation on this approach will be conducted as follows: advantages, disadvantages, and considerations.

5.1 Advantages

The advantages of the MLB are that the use of a blockchain to store a user's digital identity inherits the advantages of a blockchain. The key advantages being that the system protects the data from being altered, prevents unauthorised access, the data is stored in a decentralised manner, and most importantly transparency and security (Naik & Jenkins, 2020). The use of multiple blockchains between the user and the digital platform, allows both the isolation of the user's digital identity, and pseudo-anonymisation on the platform. This isolation prevents the user's information being compromised on the digital platform due to the information not being stored on the platform, this will also reduce the number of user information motivated attacks on the platform. Should there be an attack on the digital platform, the information stored would be stored on the internal blockchain – making the data gathered useless unless the attackers have the time and resources to decrypt the contents of the blockchain.

The use of claims to allow digital platforms to request access to information from the user, along with a time-to-live limitation on the data controls the information that the digital platform is allowed to access. This restriction prevents the digital platform from storing the user's information indefinitely and unsecured, as well as prevents the platform from distributing the data to third parties without the user's consent. The use of claim requests also makes the user aware of where their information is being used and the purpose for requesting their information, the user is then given full authority to deny or allow the request should they feel fit.

5.2 Disadvantages

As mentioned in the previous section, the MLB inherits the traditional properties of blockchain technology, this means that the MLB inherits blockchain's disadvantages as well. Due to the decentralised nature of blockchain, if either the Global Blockchain or Internal Blockchain is compromised, then the data stored in the respective blockchain will be compromised (Dai, et. al., 2017), while this will not be an issue on the digital platform's internal blockchain regarding user data, this is of concern on the Global Blockchain due to the user identities being stored on the blockchain, future work on this research will attempt to prevent this from occurring.

The energy consumption and scalability of the MLB is another aspect to consider, while this is not much of a concern on the Global Blockchain, it is a concern on the digital platform's internal blockchain, depending on the frequency of use that the digital platform has. The Global Blockchain is not of concern regarding scalability and energy consumption when generating nodes, as the number of nodes being directly proportional to the population of the digital space, i.e. one node per digital identity and only one digital identity can be associated with a person in the world. This means that on initial creation of the Global Blockchain, energy consumption will be high. However, once this process is complete consumption will be significantly reduced when a digital identity is added after this process.

5.3 Considerations

While there are advantages and disadvantages of the MLB, there are considerations that need to be accounted for regarding the authority of the system. The authority that controls the claims in the system will need to be an entity that is trusted by all parties involved, being impartial, with the integrity to assign and sign claims for digital platforms. This is to ensure that user trust is enforced between the digital platform and that the user feels safe that their information is not being used for purposes other than what it was intended for.

The second consideration regarding authority is that the global blockchain needs to be a private blockchain. This is to prevent the chain from being compromised by a syndicate of malicious nodes as well as prevent the blocks' data on the blockchain from being readily available to the public or malicious actors. A block's data can include the historical data of the chain, headers, data, and importantly peer information. By limiting the nodes to trusted entities, we can prevent peer information from being available to malicious actors, preventing attacks on other nodes in the blockchain.

An important consideration is the impact that the MLB would have on current authentication schemes as well as the architectural changes that would be needed to implement the MLB on a global scale.

6. Conclusion and Future work

This paper introduces a model of the Multi-level Blockchain (MLB) a system which consists of two blockchains which aims to ensure user privacy and prevent user information from being compromised when data breaches occur, while also preventing user data duplication between digital platforms from occurring. Consisting of a Global Blockchain containing users' digital identities, and digital platforms Internal Blockchain containing platform specific information, the MLB can achieve this aim by both securing a user's digital identity and isolating their identity from a digital platform.

The next steps pertaining to this model is to implement the model programmatically and experimenting further by simulating a working example of this model. This would be able to examine the system and bring forth any limitations or pitfalls which this system may occur that has not been discussed in this paper as well as a comparison of the MLB compared to existing authentication schemes. Future work will also research the aspects of the system mentioned in the evaluation section, bringing into consideration the disadvantages, and determining methods to minimise these disadvantages.

References

Ante, L., Fischer, C., & Strehle, E. (2022). A bibliometric review of research on digital identity: Research streams, influential works and future research paths. *Journal of Manufacturing Systems, 62*, 523 - 538. doi:https://doi.org/10.1016/j.jmsy.2022.01.005

BBC News. (2014, 09 2). *Apple confirms accounts compromised but denies security breach*. Retrieved September 19, 2023, from BBC News: https://www.bbc.com/news/technology-29039294

Chan, C., & Virkki, J. (2014, January). Perspectives for Sharing Personal Information on Online Social Networks. *Social Networking, 3*. doi:10.4236/sn.2014.31005

Dai, F., Shi, Y., Meng, N., Wei, L., & Ye, Z. (2017). From Bitcoin to cybersecurity: A comparative study of blockchain application and security issues. *2017 4th International Conference on Systems and Informatics (ICSAI)*, (pp. 975-979). doi:10.1109/ICSAI.2017.8248427

Gao, X., Yu, L., He, H., Wang, X., & Wang, Y. (2020). A research of security in website account binding. *Journal of Information Security and Applications, 51*. doi:10.1016/j.jisa.2019.102444

Hashimova, K. (2016, 07). The Role of Big Data in Internet Advertising Problem Solution. *International Journal of Education and Management Engineering, 6*, 10 - 19. doi:10.5815/ijeme.2016.04.02

Hinds, J., Williams, E. J., & Joinson, A. N. (2020). "It wouldn't happen to me": Privacy concerns and perspectives following the Cambridge Analytica scandal. *International Journal of Human-Computer Studies, 143*.

Karie, N. M., Kebande, V. R., Ikuesan, R. A., Sookhak, M., & Venter, H. S. (2020). Hardening SAML by Integrating SSO and Multi-Factor Authentication (MFA) in the Cloud. *Proceedings of the 3rd International Conference on Networking, Information Systems & Security, NISS '20*(56), 6. doi:10.1145/3386723.3387875

Marwick, A. E. (2017, 09). Scandal or sex crime? Gendered privacy and the celebrity nude photo leaks. *Ethics and Information Technology, 19*(3), 191. doi:10.1007/s10676-017-9431-7

Mohammed, Z. A., & Tejay, G. P. (2023, February). How to Compensate After a Data Breach? Investigating Compensation Types and Role of Fairness in Customer Repatronage Intentions. *SIGMIS Database, 54*, 110–127. doi:10.1145/3583581.3583588

Mohsin, A. H., Zaidan, A. A., Zaidan, B. B., Albahri, O. S., Albahri, A. S., Alsalem, M. A., & Mohammed, K. I. (2019, May). Blockchain Authentication of Network Applications: Taxonomy, Classification, Capabilities, Open Challenges, Motivations, Recommendations and Future Directions. *Comput. Stand. Interfaces, 64*, 41–60. doi:10.1016/j.csi.2018.12.002

Naik, N., & Jenkins, P. (2020). Governing Principles of Self-Sovereign Identity Applied to Blockchain Enabled Privacy Preserving Identity Management Systems. *2020 IEEE International Symposium on Systems Engineering (ISSE)*, (pp. 1-6). doi:10.1109/ISSE49799.2020.9272212

Neto, N. N., Madnick, S., Paula, A. M., & Borges, N. M. (2021, January). Developing a Global Data Breach Database and the Challenges Encountered. *J. Data and Information Quality, 13*. doi:10.1145/3439873

Pham, T.-H., Vo, Q.-H., Dao, H., & Fukuda, K. (2023). SSOLogin: A framework for automated web privacy measurement with SSO logins. *Proceedings of the 18th Asian Internet Engineering Conference, AINTEC '23*, 69–77. doi:10.1145/3630590.3630599

Sadqi, Y., Belfaik, Y., & Safi, S. (2020). Web OAuth-Based SSO Systems Security. *Proceedings of the 3rd International Conference on Networking, Information Systems \& Security, NISS '20*(69), 7. doi:10.1145/3386723.3387888

Saglam, R. B., Nurse, J. R., & Hodges, D. (2022). Personal information: Perceptions, types and evolution. *Journal of Information Security and Applications, 66*, 103163. doi:https://doi.org/10.1016/j.jisa.2022.103163

Saklikar, S., & Saha, S. (2006). User Privacy-Preserving Identity Data Dependencies. *Proceedings of the Second ACM Workshop on Digital Identity Management* (pp. 45 - 54). Alexandria, USA: Association for Computing Machinery. doi:10.1145/1179529.1179537

Shipman, F. M., & Marshall, C. C. (2020). Ownership, Privacy, and Control in the Wake of Cambridge Analytica: The Relationship between Attitudes and Awareness. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1 - 12). New York, NY, USA: Association for Computing Machinery. doi:10.1145/3313831.3376662

Takemiya, M., & Vanieiev, B. (2018). Sora Identity: Secure, Digital Identity on the Blockchain. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, *02*, pp. 582-587. doi:10.1109/COMPSAC.2018.10299

Zhang, R., Xue, R., & Liu, L. (2019, July). Security and Privacy on Blockchain. *ACM Comput. Surv., 52*. doi:10.1145/3316481

Defining the "Cyber-Pearl Harbor" - Validation of the DSLP-framework in "Offensive Cyberspace Operations Targeting Ukraine: A Cyber Pearl-Harbor"

Jonathan Lilja[1], Eleshwa Eishayea[1] and Gazmend Huskaj[1,2]
[1]Department of Computer and Systems Sciences, Kista, Sweden
[2]Geneva Centre for Security Policy, Geneva, Switzerland

jonathan.lilja@sandvik.com
eleshwa.eishayea@outlook.com
g.huskaj@gcsp.ch

Abstract: The use of cyberattacks against organizations, health care and individuals have increased along with the constant digitalisation. Nations have also fallen victim to cyberattacks, often combined with other means of war. A Cyber-Pearl Harbor (further shortened as CPH) is a term mentioned by the former United States Secretary of Defense, Mr. Leon Panetta, who described it as "combined attacks that result in human death and physical destruction and that paralyzes an entire nation". Gazmend Huskaj used Panetta's definition in order to create a framework (herein the "DSLP-framework") for classifying an event as a CPH. This study strives to see if the DSLP-framework can be validated since the term has been widely used for the last 25 years. However, a universal definition of the term seems to be missing, therefore it is not certain that the criteria presented in the framework is accurate. A qualitative case study was conducted through a literature review regarding the term CPH and semi structured interviews with three experts were done, which were later analysed through a thematic analysis. The framework was applied to three real life cases: a public health sector in Costa Rica, a TV-tower in Kyiv, Ukraine and the grocery company Coop in Sweden. The result from applying the DSLP-framework to each case was that only the case of TV-tower in Kiev, Ukraine, could be classified as a CPH. The following conclusions were drawn in this study: 1) The framework could not be validated due to lack of data, and 2) The interviewees views differed from the framework making it hard to find common ground.

Keywords: Cyber-Pearl Harbor, Cyber attacks, Cyber security, Hybrid operations, Russia-Ukraine war

1. Introduction

Attacks against computers or networks, commonly known as cyber attacks, have become increasingly prevalent in recent years (ENISA, 2022). With the advent of the Internet of Things (IoT), new and more complex forms of cyber attacks are continually emerging, particularly in warfare. Here, a blend of kinetic and cyber means is increasingly observed. The ongoing Russia-Ukraine war exemplifies this, where physical attacks using missiles and cyber attacks targeting infrastructure occur frequently. This situation raises questions about the extent of these attacks and whether they can be classified as Cyber Pearl Harbors (CPH).

Former United States Secretary of Defense, Mr. Leon Panetta, described a CPH as "an attack that would cause physical destruction and the loss of life, paralyzing and shocking the nation and creating a new, profound sense of vulnerability" (U.S. Department of Defense, 2012). He also noted that such an attack includes targeted attacks on key elements of a nation's critical infrastructure, rendering communication and military systems either unusable or severely limited (ibid.). Huskaj (2023) also refers to this definition and, through linguistic analysis, deconstructs it into components used in his framework, as depicted in **Figure 1:** 1) The attack comprises both cyber and physical elements; 2) The impact of the attack includes physical destruction and loss of human life; 3) The nation is plunged into a state of cognitive shock (cognitive effect).

Component 1 – Attack Type	Component 2 – Impact of Attack
Cyber attacks AND physical attacks	Physical destruction AND loss of life
Component 3 – Cognitive Effects	
Bring into being a deep, fresh, perceived wounding	
[PICTURE TO VISUALLY PRESENT THE PERCEIVED WOUNDING]	

Figure 1: The DSLP-framework

The purpose of this research product is to answer the research question: How can the "DSLP-Framework" in "Offensive Cyberspace Operations Targeting Ukraine: a Cyber Pearl-Harbor" (2023) be validated for the classification of cyber attacks as Cyber Pearl-Harbor?

By examining whether real-world examples that meet the DSLP requirements can be found in today's environment, a successful validation could enhance the framework's credibility. This, in turn, would facilitate its application in future cases. The validation was conducted using the following three real-life cases, all of which share the common characteristic of attacks focusing on disrupting systems and their availability: the attack on the public healthcare sector Caja Costarricense de Seguro Social in Costa Rica, the TV tower in Kyiv, Ukraine, and the grocery company Coop in Sweden.

In late May 2022, Costa Rica's public health sector, Caja Costarricense de Seguro Social (CCSS), suffered a ransomware attack. This incident resulted in 30 CCSS servers being infected and approximately 1,200 hospitals and clinics being shut down (Murillo, 2022; Córdoba & Sherman, 2022; Caja Costarricense de Seguro Social, 2022).

On March 1, 2022, Russia launched a malware attack ("DesertBlade") against a TV tower in the capital, Kyiv, aiming to destroy and silence TV media and channels in Ukraine. On the same day, two Russian missiles reportedly hit the TV tower, killing five people and injuring five others (Microsoft, 2022; Harding, 2022; Koch-Emmery & Wikén, 2022).

Beyond politics and warfare, large companies have also been targets of major cyberattacks. In Sweden, the grocery chain Coop experienced a cyber attack that crashed their cash register system due to the software company Kaseya being subjected to a ransomware attack (Toresson, 2021).

The remainder of the article is organized as follows: Section 2 describes the methods and materials; Section 3 presents the results of the study; Section 4 discusses these results; and Section 5 concludes with the conclusions and directions for future research.

2. Methods and Materials

Case studies aim to study and examine a phenomenon in depth and within its contemporary context (Johannesson & Perjons, 2014). This research strategy provides a detailed description of the research, providing a holistic perspective of the phenomenon, its processes, and its potential relationships within its context (Denscombe, 2014). Since this study seeks to investigate and validate whether a framework developed for Cyber Pearl Harbor (CPH) can be applied to three additional cases, a case study approach was deemed the most suitable research strategy. Furthermore, the flexible nature of this strategy not only facilitates the creation of new theories and/or the validation of existing frameworks but also allows for the integration of other research strategies (Denscombe, 2014). These characteristics are particularly relevant in this study, where a validation of an existing framework, published by Huskaj (2023), is undertaken. Thus, case studies are the appropriate strategy for this study.

Semi-structured interviews were conducted with the aim of collecting further data from experts, all of whom possess well-founded knowledge in the field of cybersecurity. A total of three interviews were carried out, either on-site or via Zoom. All interviewees have a background in defence, with two possessing substantial knowledge in technical means and hybrid attacks, while the third interviewee had deeper expertise in influence operations. One of the interviews was conducted in English and was intentionally not translated to retain the context that might otherwise have been lost.

To analyse the collected data and identify potential themes and patterns, the method of thematic analysis was employed. The flexibility of this method allows a researcher to identify, analyse, and report on the key themes and patterns emerging from the data (Braun & Clarke, 2006). The thematic analysis was conducted on the data collected from the literature review and the semi-structured interviews (see **Figure 2**). The review and conceptualization of the scientific research on "Cyber Pearl Harbor" are presented in **Table 1**.

Table 1: Conceptualisation derived from the literature review

Concept	Definition	References
Cyber-Pearl Harbor	A combined attack consisting of cyber and physical operations leading to loss of human life and cognitive impact on a society.	(Huskaj 2023)
Definition of the term	Researchers disagree on definitions. More research is needed to come up with a universal definition.	(Lawson & Middleton 2019; U.S. Department of Defense 2012; Smith 1998; Straub 2021; Huskaj 2023)
Probability of occurrence	There are differences of opinion on the likelihood that a Cyber Pearl Harbor would occur/has already occurred.	(Smith 1998; Straub 2021; Goldman & Warner 2017; Libicki

Concept	Definition	References
		2009; Wirtz 2018; U.S. Department of Defense 2012; Clark & Knake 2010)
Relevance of the term	Researchers disagree whether the term is still relevant.	(Kallberg 2021; Vavra 2021; Wirtz 2017)
Grey-zone	Characterised by an area between cyber warfare and political warfare that plays on a line between war and peace.	(Wither 2022; Dziwisz 2022)
Cyber disinformation	Dissemination of disinformation in the form of deep fakes, propaganda and misinformation with the aim of causing damage at a cognitive level.	(ENISA 2022; Microsoft 2022b; Whyte & Mazanec 2019)
DDoS attacks	Abbreviation for "Distributed Denial of Service". A type of cyber attack where multiple vulnerable computers are used to overload a victim's system, network or server with a large amount of traffic and make it inaccessible to users.	(ENISA 2022)
Ransomware attacks	A type of malware that encrypts a victim's files. Ransom is then used to restore access to the encrypted files. It often results in significant disruption and financial loss to victims.	(ENISA 2022)
National level	Cyber-attacks occur on a national level.	(ENISA 2022; Microsoft 2022a; Microsoft 2022b; Wither 2022; Dziwisz 2022)
Organisational level	Cyber-attacks also target businesses and the public sector, such as the healthcare sector.	(ENISA 2022; Riggi 2020; Microsoft 2022b)

2.1 Limitations of the Study

This study is subject to several limitations. Firstly, time was a significant constraint due to the project's duration of six months on a half-time basis (equivalent to 3 months full-time). This temporal limitation significantly impacted the research, as it restricted the authors to conducting only three semi-structured interviews. Consequently, this leads to the second limitation: restricted access to empirical data. A larger number of respondents would have likely yielded more empirical data, thereby potentially providing a clearer insight into how respondents perceive and interpret the concept of a “Cyber Pearl Harbor.” Finally, a limitation concerning the study's conclusions must be acknowledged: the limited scope and depth of the research preclude the drawing of comprehensive conclusions.

3. Results

In this section the results of the thematic analysis and case study are presented.

3.1 Thematic Analysis

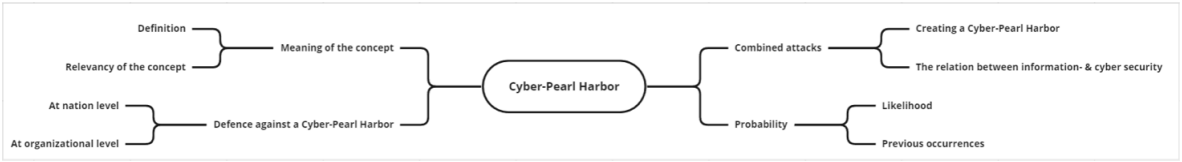


Figure 2: The thematic analysis resulted in four main themes and eight categories

3.1.1 Meaning of the concept

Meaning of concepts encompasses two categories: *Definition* and *Relevancy of the concept*. The first category addresses the disagreement over the definition of “Cyber Pearl Harbor,” highlighting the need for both a universal standard and a redefinition of the concept. In efforts to define the concept, the element of surprise is frequently mentioned, with an emphasis on the critical effects such attacks can have on society. The scope of the concept was also explored, including the aspect of utilising a combination of tools in executing a Cyber Pearl Harbor. To classify an attack as a Cyber Pearl Harbor, the involvement of multiple sectors of society or an entire country is necessary, achievable through the combination of tools and other means, as previously mentioned. The potential effects of a Cyber Pearl Harbor were also discussed, raising questions about the possibility of loss of life and examining the criteria described in Huskaj's (2023) DSLP-framework. The importance of involving relevant entities in defining the term was also underscored.

The second category involves discussions regarding the relevancy of the Cyber Pearl Harbor concept. The concept was described as outdated, needing to evolve in line with contemporary usage of the term and ongoing digitalization. The manner in which the term has changed over the years was a topic that elicited divided opinions.

3.1.2 Defence against a Cyber-Pearl Harbor

The theme of *Defence against a Cyber-Pearl Harbor* encompasses strategies at both the *national* and *organisational* levels. At the national level, the discussion focused on the implementation of an "emergency plan" that every country should follow in the event of a Cyber Pearl Harbor. The concept of cyber resilience was also mentioned in the context of building up national defence.

At the organisational level, the emphasis was placed on establishing routines and the importance of cooperation within and between different businesses. Such cooperation would contribute to the knowledge pool within an organization and serve as an opportunity to learn from and with other entities. In the event of a Cyber Pearl Harbor, these entities could assist each other, or the emergency plan mentioned previously could also function as an effective measure.

3.1.3 Combined attacks

Combined attacks were a central theme in all interviews. Interviewees shared the view that cyber-attacks, on their own, do not yet possess the scale required to create effects as described in the DSLP-framework. The importance of employing multiple means of action simultaneously was emphasised, particularly in understanding what could constitute a Cyber Pearl Harbor (CPH) in practice. This includes not only the execution of CPH but also how cyber-attacks are intertwined with influence operations and should be considered from a defence perspective, linking information and cyber security. There were also multiple mentions of the strategy of launching attacks in different locations simultaneously to achieve maximum effect. Such a strategy can confuse the adversary and create a sense of desperation, especially when the attacks target resources essential to the functioning of society.

The interviewees unanimously agreed that discussions on cybersecurity cannot exclude information security. From an offensive standpoint, this linkage is often manifested through the simultaneous use of physical and cyber-attacks.

3.1.4 Probability

In the context of the Russia-Ukraine War, discussions arose regarding the probability of a Cyber Pearl Harbor and the factors necessary to facilitate it. To estimate this probability, it is essential to consider *previous incidents*. All interviewees concurred that a Cyber Pearl Harbor has not yet occurred. This consensus is attributed to several factors, such as the insufficiency of single cyber attacks in achieving the crippling effect required by the DSLP definition without the support of combined means of action. Additionally, cyber attacks have not been extensive enough to impact an entire society. This might be due to foreign military powers allocating more resources to conventional means than to cyber, and the challenge in identifying clear instances where cyber means alone have resulted in human fatalities. It is also evident that countries frequently targeted by cyber attacks have developed cyber resilience as a defence, and adversaries have not been prepared to encounter such resistance in the cyber domain.

Opinions varied regarding the likelihood of a Cyber Pearl Harbor in the future. Interviewee 1 suggested that the probability of a nationwide attack is high, but it is improbable to paralyse a country for an extended period. Interviewee 2 questioned whether society is sufficiently uniform and integrated to be wholly impacted by a potential Cyber Pearl Harbor. Conversely, Interviewee 3 focused more on when a Cyber Pearl Harbor will occur rather than if it will occur, citing the advancement of AI and similar digital technologies as potential significant threats.

However, all interviewees emphasized that it is currently impossible to predict with certainty when a Cyber Pearl Harbor will occur. This uncertainty is partly due to the element of surprise encompassed in the DSLP-framework's definition, which can also be applied to previous war incidents that have shocked the world.

3.2 Validation of the DSLP-Framework

The validation shows that only one out of three cases could be classified as a Cyber-Pearl Harbor (*TV-tower in Kiev, Ukraine*), which is presented in the table below.

Table 2: Cases versus the DSLP-framework

Case	Criteria	
Public healthcare sector, Caja Costarricense de Seguro Social in Costa Rica	Cyber attacks	Kinetic attacks
	Physical destruction	Loss of life
	Cognitive effects	
Tv-tower in Kiev, Ukraine	Cyber attacks	Kinetic attacks
	Physical destruction	Loss of life
	Cognitive effects	
Grocery company Coop in Sweden	Cyber attacks	Kinetic attacks
	Physical destruction	Loss of life
	Cognitive effects	

3.2.1 Public healthcare sector Caja Costarricense de Seguro Social in Costa Rica

The cyber-attack against the CCSS in Costa Rica was a ransomware attack in which approximately 30 servers were directly infected by the virus (Córdoba 2022). No physical attacks were reported in connection with this cyber-attack. However, a significant consequence of the cyber-attack was the closure of around 1,200 hospitals and clinics (Murillo 2022). Several patients were affected by the cyber-attack and its repercussions, but fortunately, no human deaths were reported. One of the cognitive effects of the cyber-attack included the disruption in reporting and accounting for COVID-19 related data (Córdoba 2022). Additionally, patients were informed that they could not receive treatment during this period, and there were reports of surgeries being delayed (Caja Costarricense de Seguro Social 2022).

In summary, the cyber-attack against the CCSS healthcare sector in Costa Rica does not qualify as a Cyber Pearl Harbor under the DSLP-framework. Despite targeting a sensitive sector where human casualties could easily occur, the attack did not coincide with any kinetic attack, nor were there any reported loss of life.

3.2.2 TV-tower in Kiev, Ukraine

The attack on the TV tower in Kyiv was a combined assault employing both physical means, in the form of missiles, and cyber means, through the “DesertBlade” trojan (Microsoft, 2022). This attack resulted in the loss of five lives and damage to hardware, leading to the disruption of broadcasting on the TV company's channels (Harding, 2022). Beyond instilling fear in Kyiv, where the attack occurred, the objective was to crush the hope and resistance of the Ukrainian people and their army. However, the (former) Minister of Defence of Ukraine expressed hopeful words, asserting that victory was the only option and that Ukraine would never surrender (ibid.). This suggests that the cognitive effects Russia intended to create were not entirely successful. Nevertheless, the event had a nationwide impact, affecting the entirety of Ukraine.

The attack on the TV tower in Kyiv can be classified as a Cyber Pearl Harbor under the DSLP-framework. The assault was a combined operation with different means of action that resulted in human casualties and had clear cognitive effects on the entire country. The only discrepancy lies in the fact that the cognitive effects did not completely paralyse Ukraine, as the nation continued to defend itself and encouraged its population to maintain hope.

3.2.3 Grocery company Coop in Sweden

The attack against Coop was executed solely through cyber means, specifically using ransomware (Toresson 2022). The physical effects of the attack included financial losses for the company, the closure of the majority of its stores, and a paralysed web shop (Truesec n.d.). The event's major cognitive impact was the increased concern among other companies about similar incidents, and the state of shock experienced by Coop during the attack (Stenberg 2022). In summary, although the attack occurred on a nationwide level for the company, it did not result in a societal paralysis. This is because Coop did not hold a monopoly on food sales, allowing

society to continue functioning relatively normally, albeit with Coop customers having to shop at other stores. The attack was surprising to society, but it was not paralysing. Moreover, it was purely a cyber-attack without any kinetic component, and no human deaths resulted from it. With these factors in mind, the Coop attack cannot be classified as a Cyber Pearl Harbor under the DSLP-framework.

4. Discussion

Panetta's definition of a Cyber Pearl Harbor describes an attack aimed at causing physical destruction, loss of life, and paralysing and shocking a nation (U.S. Department of Defense, 2012). However, this definition does not align with those shared by the interviewees, highlighting the current lack of a universal standard for the concept. Discrepancies were noted either in specific aspects of the concept, such as the element of surprise or the combination of several tools, or in the overall definition itself, such as questioning whether a Cyber Pearl Harbor could lead to loss of life. Due to this inability to agree on a standard definition, the probability of a Cyber Pearl Harbor was also dismissed entirely, contradicting the conclusions drawn by Huskaj (2023) in his article. This divergence could be a result of the absence of a universal definition and uncertainty about what exactly is required for an attack to be classified as a Cyber Pearl Harbor. This ambiguity could pose significant challenges in establishing prevention strategies within nations and organisations. Without an international standard to follow, a nation or organisation may expose itself to greater vulnerability and weaknesses than if such a standard existed.

Another key point raised was the scope of the concept, emphasising the importance of a Cyber Pearl Harbor being comprehensive, i.e., involving multiple entities. This aspect aligns with Panetta's definition, where a Cyber Pearl Harbor includes attacks against national infrastructure with the objective of significantly limiting or completely eliminating it.

The theme of combined tools was a recurring topic in all interviews, along with the consensus that a Cyber Pearl Harbor cannot currently be executed using cyber means alone. The interviewees noted that for an attack to be classified as a Cyber Pearl Harbor, it must result in a major impact, a result that the combination of tools can ensure. Although the use of combined tools is not explicitly mentioned in the DSLP-framework, it is hinted at in the criterion of “Cyber attacks & Kinetic attacks.” This criterion implies that a combination is occurring, as it involves the use of both cyber and kinetic attacks. It is noteworthy that the interviewees did not recognise this subtle connection, further exemplifying the confusion surrounding the concept of Cyber Pearl Harbor and how it persists due to the lack of a universal definition. Similar to the challenges in defining the overall concept of Cyber Pearl Harbor, additional uncertainties exist regarding what exactly constitutes a cyber attack, what constitutes a kinetic attack, and what actions are necessary to achieve these.

The respondents disagreed with the framework regarding whether a Cyber Pearl Harbor (CPH) has already occurred. In Huskaj's article (2023), where the framework was introduced, he suggests that this disagreement may stem from individual perceptions of what constitutes a CPH, as well as a narrow-minded approach among some cybersecurity experts. This conclusion is also supported by the literature review and interviews conducted in this study. One reason experts claim that a CPH has not yet occurred could be due to the rapid development of technology and digital means, which may have created a distorted perception of the power of combined attacks and the significance of cyber means, including influence operations such as deep fakes and disinformation.

During the interviews, estimates of the likelihood of a CPH were also discussed. Two out of three interviewees believe that the probability of a single attack reaching the level of a CPH is low, but they emphasise the uncertainty surrounding this assessment. The third interviewee anticipates a CPH occurring in the Russia-Ukraine war, viewing it as a potential decisive tool in the conflict. This contradicts the framework, which used cases from the Russia-Ukraine war and classified the event as a CPH. Generally, however, this interviewee perceives a high probability of a CPH, suggesting that it may already exist within our systems, yet to be executed. The interviewee argues that the constant digitalisation of society, increased use of IoT, and development of AI are creating vulnerabilities in nations for cyber warfare, necessitating more resources for cyber defence. Cyber-attacks may have evolved since 2012, when Panetta highlighted the issue, to a point where kinetic attacks are no longer required to cause human death, infrastructure failure, and shock. However, this evolution has not yet been widely recognised. The framework is based on Panetta's speech, which served as a warning about the threat posed by a CPH and the importance for nations to take it seriously and build defences against it.

Although the cases presented in this study did not meet the set criteria for a CPH, there is no indication that a future occurrence is impossible. This is particularly relevant in the context of cyberattacks against healthcare, a sector that has seen an increase in all forms of cyberattacks over the years. While the sector may not yet have fallen victim to a cyberattack at the level of a CPH, considering its crucial role in society and its high sensitivity, such an attack could very well occur (or may have already occurred without being noticed).

5. Conclusions

The results suggest that the current capabilities of cyber operations are not sufficient to create an event comparable to what might be classified as a Cyber Pearl Harbor (CPH); a combination of means is required, as the framework supports. While the framework posits that a CPH has already occurred, this case study's findings do not corroborate that assertion. The DSLP-framework does not explicitly state the likelihood of a CPH occurring. However, Huskaj's (2023) article, which identifies two instances of a CPH, allows for the assumption that the probability is relatively high. This study's results are mixed, with some aspects supporting this assumption and others contradicting it, leading to the conclusion that the probability of a CPH occurring remains indeterminate.

Currently, validating the framework is not feasible due to the insufficient data available to fully substantiate it. The majority of the existing data does not support the framework. Although there may be similar perceptions of the concept among respondents, the differences are too significant to form a consensus. Therefore, it is essential to involve a broader range of participants to collect more data, contributing to the development of a universal definition of the term.

Future research should aim to include a larger and more diverse group of respondents to gain a more comprehensive understanding of the research community's views, particularly regarding whether the majority align with Huskaj's definition of a CPH. Additionally, analysing suitable defence measures for attacks of this magnitude would contribute to enhancing cyber resilience. Further in-depth research into the cognitive effects of various attacks could also be valuable, potentially aiding in the future validation of the framework, especially in the context of a major hybrid operation.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.

Caja Costarricense de Seguro Social. (2022). #Importante El Hospital de Niños urge comunicarse con familiares de pacientes en espera de cirugía ambulatoria referidos al área de salud Tibás [...] [Facebook post]. Facebook. https://www.facebook.com/ccssdecostarica/posts/5552368771453467

Clarke, R. A., & Knake, R. (2010). *Cyber War: The Next Threat to National Security and What to Do About It*. HarperCollins.

Córdoba, J. (2022). Costa Rica public health system targeted by ransomware. *Associated Press News*. https://apnews.com/article/russia-ukraine-covid-politics-technology-health-0e24e6644b09e2737af96814635fcd22

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative* (4th ed.). Pearson.

Denscombe, M. (2014). *The Good Research Guide: For Small-Scale Social Research Projects* (5th ed.). Open University Press.

Dziwisz, D. (2022). Cyber Pearl Harbor is not coming: US politics between war and peace. *Understanding Contemporary Security*, 19(4), 95-109. https://doi.org/10.12797/Politeja.19.2022.79.07

ENISA. (2022). ENISA Threat Landscape 2022. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022

Goldman, E. O., & Warner, M. (2017). Why a digital Pearl Harbor makes sense ... and is possible. In G. Perkovich & A. E. Levite (Eds.), *Understanding Cyber Conflict: Fourteen* Analogies (pp. 147-157). Georgetown University Press.

Harding, L. (2022). Ukraine says Russia targeting civilians as missiles hit Kyiv TV tower. *The Guardian*. https://www.theguardian.com/world/2022/mar/01/ukraine-russia-civilians-missiles-kyiv-tv-tower

Huskaj, G. (2023). Opérations offensives dans le cyberespace ciblant l'Ukraine: Un cyber Pearl Harbor? *Revue Militaire Suisse, Artikel Numéro Thématique - Ukraine - 2023*, 42-44.

Johannesson, P., & Perjons, E. (2014). *An Introduction to Design Science.* Springer International Publishing.

Kallberg, J. (2021). Bye bye, cyber Pearl Harbor. *CyberWire*. https://www.researchgate.net/publication/351980025_Bye_bye_cyber_Pearl_Harbor

Koch-Emmery, L., & Wikén, E. (2022). Attack mot tv-torn i Kiev – "Tystar medier som når ut längst". *SVT*. https://www.svt.se/nyheter/utrikes/attack-mot-tv-torn-i-kiev-tystar-medier-som-nar-ut-langst

Lawson, S., & Middleton, M. K. (2019). Cyber Pearl Harbor: Analogy, fear, and the framing of cyber security threats in the United States, 1991-2016. *First Monday, 24*(3). https://doi.org/10.5210/fm.v24i3.9623

Libicki, M. C. (2009). *Cyberdeterrence and Cyberwar*. RAND Corporation.

Microsoft. (2022a). Defending Ukraine: Early lessons from the cyber war. https://blogs.microsoft.com/on-the-issues/2022/06/22/defending-ukraine-early-lessons-from-the-cyber-war/

Microsoft. (2022b). An overview of Russia's cyberattack activity in Ukraine. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RE4Vwwd

Murillo, A. (2022). Latest cyberattack in Costa Rica targets hospital system. *Reuters*. https://www.reuters.com/world/americas/latest-cyberattack-costa-rica-targets-hospital-system-2022-05-31/

Riggi, J. (2020). Ransomware attacks on hospitals have changed. American Hospital Association. https://www.aha.org/center/cybersecurity-and-risk-advisory-services/ransomware-attacks-hospitals-have-changed

Smith, G. (1998). An electronic Pearl Harbor? Not likely. *Issues in Science and Technology, 15*(1).

Stenberg, J. (2022). Experter: Därför är Coop-attacken bara början. *Tidningen Näringslivet*. https://www.tn.se/naringsliv/18119/experter-darfor-ar-coop-attacken-bara-borjan/

Straub, J. (2021). Defining, evaluating, preparing for, and responding to a cyber Pearl Harbor. *Technology in Society, 65*, Article 101599. https://doi.org/10.1016/j.techsoc.2021.101599

Toresson, J. (2021). It-attacken mot Coop – detta har hänt. *SVT*. https://www.svt.se/nyheter/inrikes/it-attacken-mot-coop-detta-har-hant

Truesec. (n.d.). Back in business after the largest ransomware attack of all time. https://www.truesec.com/why-truesec/cases/coop-back-in-business-after-the-largest-ransomware-attack-of-all-time

U.S. Department of Defense. (2012). Remarks by Secretary Panetta on cybersecurity to the Business Executives for National Security, New York City. https://content.govdelivery.com/accounts/USDOD/bulletins/571813

Vavra, S. (2021). What gets lost in "cyber Pearl Harbor"-style rhetoric. *CYBERSCOOP*. https://cyberscoop.com/cyber-pearl-harbor-911-cyberwar-hacking-leon-panetta-ciaran-martin/

Wirtz, J. J. (2018). The cyber Pearl Harbor redux: Helpful analogy or cyber hype? *Intelligence and National Security, 33(*5), 771-773. https://doi.org/10.1080/02684527.2018.1460087

Wirtz, J. J. (2017). The cyber Pearl Harbor. *Intelligence and National Security, 32*(6), 758–767. https://doi.org/10.1080/02684527.2017.1294379

Wither, J. K. (2022). Trends in warfare. In K. R. Lester (Ed.), *Encyclopedia of Violence, Peace, & Conflict* (3rd ed., pp. 241-253). Academic Press.

Whyte, C., & Mazanec, B. (2019). *Understanding Cyber Warfare: Politics, Policy and Strategy*. Routledge.

Assessing the Effectiveness of ADS-B Mitigations

Heinke Lubbe, Rudi Serfontein and Marijke Coetzee
School of Computer Science and Information Systems, North-West University, Potchefstroom, South Africa

lubbeheinke7@gmail.com
Rudi.Serfontein@nwu.ac.za
Marijke.Coetzee@nwu.ac.za

Abstract: The rise in aerial traffic necessitates aircraft localisation methods that go beyond radar technology's built-in capabilities. The Automated Dependent Surveillance-Broadcast (ADS-B) system is a novel aircraft localisation method that promises to provide the necessary precision to handle the current air traffic surge. The Federal Aviation Administration has, therefore, enforced ADS-B's deployment. However, the architecture of the ADS-B system holds several vulnerabilities. Most of these vulnerabilities are because ADS-B is designed to rely on wireless networks. This paper provides an in-depth analysis of the ADS-B threat landscape and potential mitigations to better understand their distinct characteristics and impact on the ADS-B system. Addressing these security concerns is imperative to ensure ADS-B systems' robustness and trustworthiness and safeguard the aviation industry from potential cyber threats. The paper concludes with a critical review of how well the proposed mitigations address the identified security threats.

Keywords: Automated Dependent Surveillance-Broadcast (ADS-B), Threats, Vulnerabilities, Mitigations, Critical Infrastructure

1. Introduction

The Automated Dependent Surveillance-Broadcast (ADS-B) system is a novel air traffic surveillance technology employed by aircraft and air traffic controllers. ADS-B allows aircraft to autonomously determine their position using a Global Positioning System (GPS) or Global Navigation Satellite System (GNSS). The ADS-B avionics then combines the data with data gathered from other aircraft systems, such as the Flight Management System (FMS), altimeter, and Traffic Collision Avoidance System (TCAS) units to construct a set of information for the aircraft (Mirzaei et al, 2019). The aircraft automatically broadcasts this information once every second to an ADS-B ground station. Ground stations send the information to an Air Traffic Controller (ATC) for reliable aircraft monitoring (Hasin et al, 2021), resulting in enhanced situational awareness and improved safety and efficiency of air traffic management. Aircraft can also communicate with one another to determine precise positions. In this regard, ADS-B IN and ADS-B OUT are the two functions that make up ADS-B (Hasin et al, 2021). ADS-B OUT is used to broadcast ADS-B data while ADS-B IN is used to receive any ADS-B broadcasted data. Figure 1 shows how ground control stations and aircraft use ADS-B IN and ADS-B OUT to exchange information.

In Figure 1, the transmitting aircraft at the top left corner uses ADS-B OUT (1) to broadcast its location, identification, velocity, and other GNSS related data through a data link (Khandker et al, 2021). There are primarily two types of data links for ADS-B data transmission: the 1090ES and UAT978 (Blåberg et al, 2020). However, 1090ES is the most widely used of these two (Khandker et al, 2022a) and operates on a frequency of 1090 MHz. Ground stations, depicted at the bottom left corner, can receive the broadcasted information using ADS-B IN (4), after which it can process the received information and then broadcast it to all other nearby aircraft using ADS-B OUT (3). Receiving aircraft equipped with ADS-B IN (2), depicted at the top left corner, can immediately receive, process, and display ADS-B signals from all other ADS-B OUT-equipped aircraft and ground stations (Khandker et al, 2021).

This research aims to unveil the ADS-B threat landscape, potential mitigation strategies, and underlying challenges. The paper contributes by not only classifying the threats according to the first two layers of the OSI model but also by critically reviewing how well the potential mitigation addresses the identified threats. In this regard, the paper identifies key areas that need to be addressed.

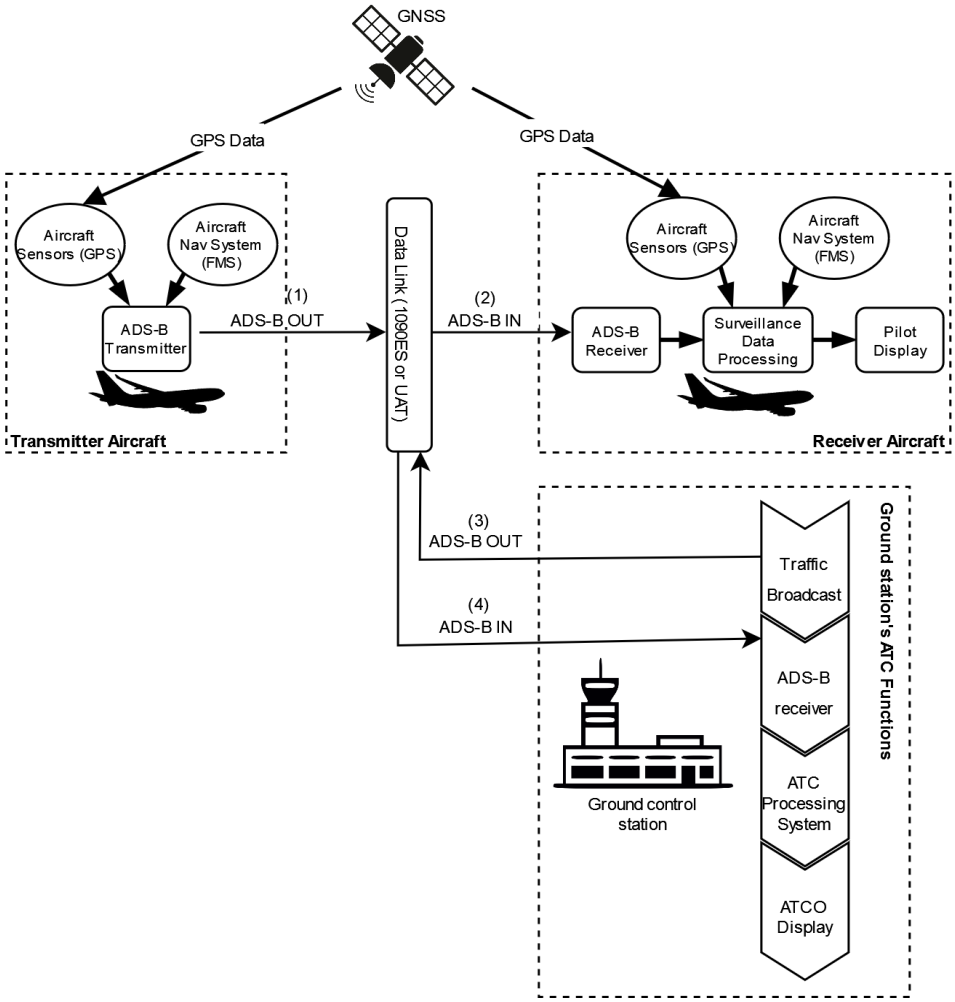


Figure 1: The ADS-B system

The following section delves into a threat analysis of the ADS-B technology to provide a better understanding of its distinct characteristics, potential ramifications, and the extent of its impact on the ADS-B system.

2. ADS-B Threats

Most ADS-B vulnerabilities result from the nature of the ADS-B architecture, as ADS-B is designed to rely on wireless networks. ADS-B's dependence on wireless networks, GPS, and low-power signals renders ADS-B more vulnerable to security vulnerabilities than radar systems (Li & Kamal, 2011). According to Manesh and Kaabouch (2017), these security vulnerabilities include message interception, injection, modification, and deletion. Figure 2 provides a classification of attacks according to the OSI model's physical and data link layers. While threats can potentially exist at higher layers, they are often less impactful to the core functionality of ADS-B compared to threats at the two lowest layers, which are critical to the system's operation. The threats are discussed in the order in which they appear in Figure 2.

2.1 Physical Layer

2.1.1 Flooding

ADS-B receivers (ASD-B IN) are designed to be sensitive. While this does allow ADS-B receivers to detect weak signals, it poses a security concern. Due to this phenomenon, ADS-B receivers are generally prone to flooding attacks whereby an attacker can flood the screen of an aircraft, or Air Traffic Control, with fake airplanes (Naganawa & Miyazaki, 2018). Flooding attacks may disrupt or overwhelm regular air traffic monitoring systems with a large volume of fabricated messages. According to Khandker et al (2021), these attacks do not require high-end equipment. However, flooding requires a high level of bandwidth in the spectrum in accordance with

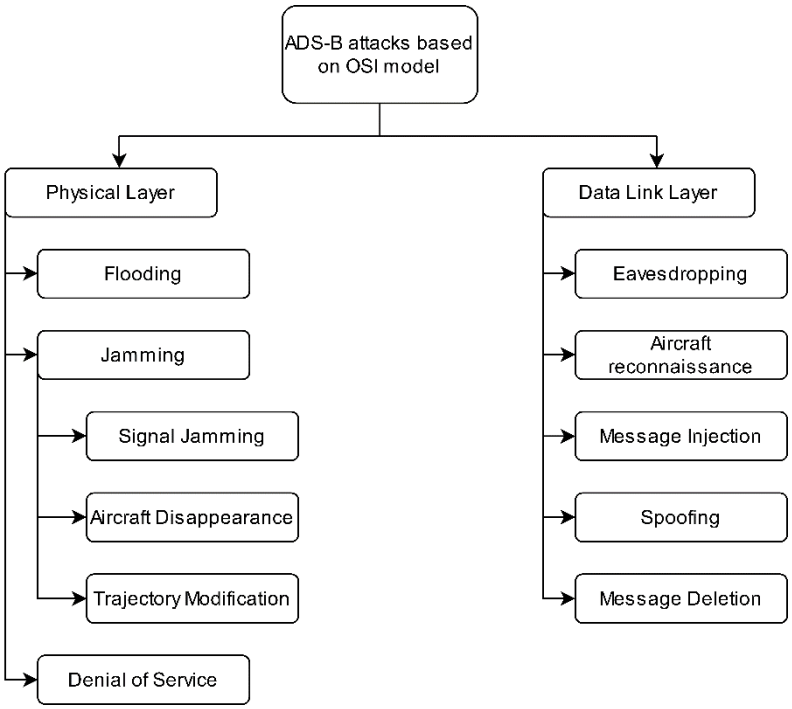


Figure 2: ADS-B Attacks categorised by OSI Model layers

It is impossible for the system to pinpoint the location of a potential attacker since ADS-B receivers have omnidirectional antennae that cannot distinguish the direction of the information they receive (Schäfer et al, 2013).

2.1.2 Jamming

According to Khandker et al (2021), an attacker can use a powerful RF signal to cause destructive interference. In this case, the attacker can deny service to wireless nodes within the interference range by overwhelming the system's spectrum. This type of attack is generally referred to as jamming.

Khandker et al (2021) outline several types of ADS-B attacks that utilise this technique:

- **Signal Jamming**: During this attack, an attacker would use a high-power RF noise transmission to block ADS-B traffic channels.
- **Aircraft Disappearance**: An attacker blocks the targeted aircraft's signal through destructive signal interference.
- **Trajectory Modification**: An attacker replaces a part of, or even the entire message, of a low-power signal using a high-power signal. The high-power signal superimposes the fake information by flipping the bits in the real message (Wu et al, 2020).

A jamming attack can either be carried out against the ground station or an aircraft (McCallie et al, 2011). When carried out against a specific ground control station, an attacker would use a low-power jamming device to overwhelm all incoming aircraft transmissions. Of importance to note is that the impact is limited to a confined area due to the jamming signal's range and proximity to the ground station. Alternatively, an attacker could perform an attack against an aircraft. An attack against an aircraft requires a high-power jamming device, as it is very challenging to get in close contact with an aircraft that is moving or flying. An attacker would consequently be more likely to launch the attack from inside an aeroplane.

2.1.3 Denial of Service (DoS)

According to Mirzaei et al (2019), no physical barrier controls access to the data link. An attacker can, therefore, send data at the same 1090MHz radio frequency to real communicating entities. ADS-B receivers can easily detect these signals since they are designed to be inherently sensitive. An attacker can consequently send a large number of fictitious signals in the form of white noise to a particular aircraft. Should the number of signals exceed the ADS-B IN capacity, the ADS-B might become unresponsive or deliver incorrect information (Khandker et al, 2021).

2.2 Data Link Layer

2.2.1 Eavesdropping and aircraft reconnaissance

The ADS-B service relies on radio communication that is not encrypted. Unencrypted radio transmissions give rise to eavesdropping as yet another security concern (Blåberg et al, 2020). According to Khandker et al (2021), unencrypted ADS-B broadcasts are accessible to everyone, making aircraft reconnaissance possible. To perform this attack, one can acquire any commercially available ADS-B receiver to intercept ADS-B OUT messages (Viveros, 2016). Wu et al (2020) extend this statement by pointing out that one can use a mobile app or public website to see real-time ADS-B information. Viveros (2016) names several public websites from which real-time ADS-B information can be obtained. The websites include www.flightradar24.com, www.radarvirtuel.com, and www.flightaware.com. The ADS-B information is accessible in real-time since the ADS-B allows aircraft to periodically broadcast information, in the form of plaintext, on a fixed frequency.

2.2.2 Message injection

The ADS-B network does not incorporate authentication in the data link layer (Manesh and Kaabouch, 2017). It is, therefore, possible to inject pseudo-legitimate messages into air traffic communications, referred to as message injection. Khandker et al (2021) state that attackers can launch such attacks using a transmission-enabled SDR-coupled device with power amplifiers. During message injection attacks, attackers construct and broadcast fake ADS-B messages with the same, but less realistic, data as real ADS-B messages, such as velocity, position, and identification number. Aircraft and Air Traffic Control systems in the vicinity receive these ADS-B messages using ADS-B IN and process the messages. As a result, a fake aircraft starts to appear among the real aircraft in the network. The goal of these types of attacks is to inject false information and disrupt the system (Khandker et al, 2021).

2.2.3 Spoofing

The ADS-B service cannot block a malicious transmission source since it relies on radio communication. In other words, the ADS-B service does not implement any form of authentication, which enables an attacker to perform a spoofing attack. These attacks are also called replay attacks and ghost aircraft injection attacks. The primary goal is to confuse the ATC system into recognising ghost planes as real planes (Ying et al, 2019). An attacker can introduce a ghost aircraft on the ATC screens by transmitting a fake signal that matches the required messaging protocol (Khandker et al, 2021). Or, more simply put, the signal must have the same attributes as an authentic ADS-B message (Wu et al, 2020). Ying et al (2019) classify spoofing attacks into the following two categories: aircraft-based and ground-based.

- **Aircraft-based:** In aircraft-based, an attacker operating from an aircraft alters the ICAO address in the ADS-B messages sent by the airborne ADS-B transponder, posing as a known or trusted aircraft to avoid detection.
- **Ground-based:** With ground-based, a ground-based attacker uses an SDR to transmit newly created and properly modulated fake messages or to retransmit previously recorded messages.

Spoofing attacks are, to a large extent, the same as message injection attacks. The only difference lies in the attack's nature and the attacker's intent. Spoofing aims to impersonate legitimate aircraft and transmit false ADS-B signals that appear authentic. Message injection, on the other hand, does not involve impersonation. Instead, the attacker injects false or unauthorised messages directly into the system without necessarily pretending to be a specific aircraft.

2.2.4 Message deletion

The fact that the ADS-B service relies on an unencrypted communication channel and cannot block a malicious transmission source allows an attacker to exploit ADS-B's CRC error handling. Attackers can exploit these 24 bits through constructive interference (Manesh & Kaabouch, 2017). The ADS-B protocol messages include 24 bits of parity (CRC error handling), which can fix up to five bits of errors. Any message with more than five incorrect bits is regarded as corrupted and dropped.

Alternatively, in destructive interference, the attacker creates a time-synchronised signal that inverses the ADS-B signal and reduces or destroys the ADS-B message (Mirzaei et al, 2019). The result is that the ADS-B signal gets destroyed while in transit. However, this approach requires the attacker to eavesdrop on the 1090 MHz extended squitter channel and create a destructive interference with the time-synchronised message that the attacker wants to delete before the message reaches the ground station (Schäfer et al, 2013). The complexity of

time synchronisation makes this kind of message deletion attack more challenging to carry out and less effective (Manesh & Kaabouch, 2017).

It becomes clear that ADS-B holds several security concerns. Addressing these security concerns is imperative to ensure ADS-B systems' robustness and trustworthiness and safeguard the aviation industry from potential cyber threats. As such, the following section discusses potential mitigations for the above-identified security concerns.

3. Mitigations

The following section provides a comprehensive analysis of the literature on potential solutions for safeguarding ADS-B systems against the threats identified in Section 2. This section aims to offer insights into techniques and strategies for enhancing the integrity and resilience of ADS-B networks.

Over the past decade, numerous studies have proposed methods for securing ADS-B communications. According to both Wu et al (2020) and Khandker et al (2021), ADS-B security solutions can be divided into two main categories: secure broadcast authentication and secure location verification. Wu et al (2020) provide an ontological tree in Figure 3 to graphically depict the security methods for ADS-B.

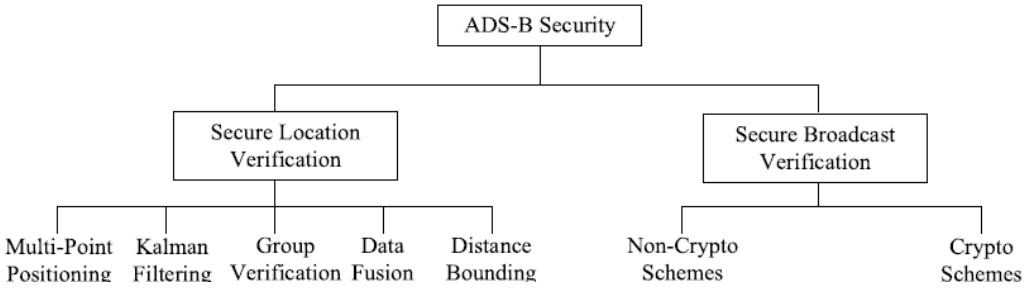


Figure 3: Classification of ADS-B mitigations (Wu et al, 2020)

The following sections discuss each category and its associated methods independently. An in-depth discussion follows the analysis of these methods.

3.1 Secure Location Verification

Secure location verification aims to determine how authentic an aircraft's position or that of another communication participant is (Wu et al, 2020). Manesh and Kaabouch (2017) state that secure location verification can be divided into two categories: secure location determination and in-region verification. Secure location determination solutions find the sender's physical location information and use that information to cross-validate the message's legitimacy (Wu et al, 2020). Conversely, in-region verification uses an estimation algorithm to calculate the likelihood that the reported position is true. In other words, in-region verification tries to assess the reliability of a stated location in an ADS-B message that has been received (Manesh & Kaabouch, 2017). The techniques used in secure location verification are analysed in the order presented in Figure 3.

3.1.1 Multi-point positioning

In multi-point positioning, an aircraft broadcasts signals to several ground stations. Each ground station receives an identical signal with a different time stamp. A hyperbolic equation calculates the aircraft's position based on the delay in the signal received (Wu et al, 2020).

3.1.2 Kalman filtering

Kalman filtering, also known as linear quadratic estimation, determines unknown variables to produce estimates that are likely more correct than estimates based solely on observations of individual measurements. The algorithm consists of three steps: Prediction, observation, and update (Manesh & Kaabouch, 2017). The system's state variables and any resulting uncertainties are anticipated in the prediction process. The forecasts rely on the system's present condition, state transitions from earlier states to its current state, and system inputs. The algorithm adjusts its estimates in the observation phase by computing the error as the difference between the current measured state measurements and the projected values derived in the previous step. The weights are updated and distributed among the estimates in the update step, with greater weights going to the estimates that have a higher probability (Bishop & Welch, 2001). The Kalman filtering can produce ideal and accurate flight status values, clear incorrect data, smooth all missing data, and filter signal noise and instability. It primarily serves the ground system's ADS-B validation needs (Wu et al, 2020).

3.1.3 Group verification

Group verification primarily uses multi-alteration technology amongst group members to confirm the location declaration information of the non-group members, hence ensuring the ADS-B IN communication security. Before joining a group, four or more aircraft must authenticate one another to establish trusted actors (Manesh & Kaabouch, 2017). The air position of non-group members can then be determined using multi-alteration techniques based on signal strength or signal arrival time differential. The aircraft in the opposite group is alerted, and the proper course of action is implemented as soon as a false location report is discovered (Wu et al, 2020).

3.1.4 Data fusion

With data fusion, data from various sources are combined and correlated to produce more accurate and trustworthy results than those from a single source (Wu et al, 2020). This technique can help mitigate attacks such as malicious data injection.

3.1.5 Distance bouncing

Distance bouncing is one of the in-region location verification methods (Manesh & Kaabouch, 2017). The basis of this method is that a signal's transmission speed (speed of an electromagnetic wave) is equal to the speed of light. Any signal delay means the message has been spoofed (Wu et al, 2020). Consider a transmitting aircraft A and a receiving aircraft B. Aircraft A becomes the prover node in distance bouncing and Aircraft B the verifier node. Aircraft B typically sends a message to Aircraft A, to which Aircraft A must respond to prove that it is in range of Aircraft B. For this reason, aircraft A is referred to as the prover node. Aircraft B, the verifier node, then considers the messages' round-trip times plus some processing time to compute the distance between the two aircraft. This information can be used to verify the correctness of the claimed location of a node (Manesh & Kaabouch, 2017).

3.1.6 Traffic modelling

Modelling is done using data mining techniques to spot inaccurate location data or other malicious behaviour. For instance, as the received signal strength is inversely proportional to the distance, one can determine whether the position data is accurate (Wu et al, 2020).

This section analysed some of the Secure Location Verification methods that are aimed at mitigating several prominent ADS-B threats. The next section examines Secure Broadcast Authentication techniques shown in Figure 3.

3.2 Secure Broadcast Authentication

Secure broadcast authentication has the security requirement that no attacker may fake a legitimate broadcast. Secure broadcast authentication seeks to offer ADS-B with an authentication method that still ensures the preservation of ADS-B's openness (Wu et al, 2020). However, Manesh and Kaabouch (2017) state that ADS-B message authentication is more challenging than for a standard point-to-point network due to the lack of two-way communications and reliable transmission between ADS-B communicating entities. There are two types of secure broadcast authentication: crypto and non-crypto schemes.

3.2.1 Crypto schemes

Symmetric and asymmetric approaches for message authentication form part of cryptographic schemes to authenticate ADS-B messages. However, the application of symmetric and asymmetric systems has several challenges due to the features of ADS-B broadcast. This seems to be particularly true concerning key management and distribution. Additionally, the frequencies designated for ADS-B technologies - 978 MHz for UAT and 1090 MHz for 1090ES - cannot support a significant number of nodes when channel interference is considerable (Manesh & Kaabouch, 2017). Cryptographic techniques that lengthen ADS-B messages result in more significant interference and constrained capacity.

Asymmetric encryption, also known as Public Key Infrastructure (PKI), is a scalable and systematic way to distribute keys. Asymmetric encryption, as an ADS-B solution, has two significant drawbacks. First, the length of sent ADS-B signals increases because current asymmetric systems lack compact encryption structures. Second, each receiver needs to receive distinct encrypted ADS-B messages (Manesh & Kaabouch, 2017). Furthermore, PKI must maintain and store certificates to assure the validity and authenticity of the public key. The revocation, updating, and storage operations involving the certificates necessary for the functioning of PKI setups will

become increasingly frequent as the number of planes rises, which is expensive and less scalable (Wu et al, 2020).

The Aircraft Address Message Authentication Code (AA-MAC) is an alternative cryptographic security solution. This specific security solution performs message integrity using a common hash technique, such as MD5 or SHA, and a private authentication key on the current aircraft address (AA) (Giannatto Jr, 2015). As such, an ADS-B message's current aircraft address (AA) field is replaced with the newly calculated AA-MAC. The rest of the message remains the same. This is required for the AA-MAC message source integrity scheme to work correctly. As with public key infrastructure, secret key distribution is also one of the difficulties that AA-MAC faces.

3.2.2 Non-crypto schemes

The above-discussed cryptographic solutions are highly dependent on key distribution and management. This requirement makes cryptographic solutions incompatible with the existing ADSB infrastructure and challenging to employ (Manesh & Kaabouch, 2017). Non-cryptographic network security strategies, however, avoid the issue of key management or distribution by primarily using fingerprinting technology and spread spectrum.

Fingerprinting, as a non-cryptographic approach to authentication and identification, is based on software or hardware errors or features of frequency channels that are difficult to replicate. By utilising fingerprinting, network intrusions can be recognised using the distinctive qualities of genuine network nodes (Zeng et al, 2010). Zeng et al (2010) outline three methods, namely software-based fingerprinting, hardware-based fingerprinting, and channel-based fingerprinting, that can be used to improve or even supplant conventional encryption.

- Software-based fingerprinting is based on the distinctive characteristics of networking software that runs on hardware from a specific vendor. These variations can be grouped and used to recognise network devices.
- Hardware-based fingerprinting rely on variations in network hardware. Hardware-based fingerprinting methods, such as radiometric fingerprinting, can take advantage of these variations in the modulation of radio waves to identify devices.
- Location-based fingerprinting use of the inherent qualities of communication channels. It has been used to replace conventional authentication and verification approaches with procedures that rely on signal intensity, channel impulse response, and carrier phase (Manesh & Kaabouch, 2017).

Each method has specific benefits and drawbacks. Some methods require considerable infrastructure, while others are relatively practical (Khandker et al, 2021).

4. ADS-B Countermeasure Gaps

The following section critically reviews the solutions discussed in this paper. In Table 1, ADS-B countermeasures are compared regarding their ability to address security threats. Jamming attacks remain the biggest concern on the physical layer, with multi-point positioning currently being the only solution. As for the data link layer, eavesdropping and aircraft reconnaissance remains a major concern. The Public Key Infrastructure (PKI) solution is the only solution to address these concerns. In fact, the PKI solution handles most security concerns despite being the most challenging and expensive. However, spoofing attacks are fully covered, followed by flooding and message deletion attacks. It should also be noted that the majority of the other defences, such as distance bounding and Kalman filtering, only address a specific attack approach and are ineffective as stand-alone remedies.

The ADS-B countermeasures are compared in Table 2 to the security services. For instance, the public key infrastructure solution guarantees data integrity, location integrity, authentication, and data confidentiality, but it cannot provide ADS-B network availability. Data fusion can ensure availability, authenticity, and location fidelity. However, the integrity and confidentiality of ADS-B transmissions are not guaranteed with data fusion.

Table 1: Comparison of ADS-B countermeasures to threats addressed

	Method	Flooding	Jamming	Denial of Service	Eavesdropping	Aircraft reconnaissance	Spoofing	Message Deletion
Secure location verification	Multi-point positioning	No	Yes	Yes	No	No	Yes	Yes
	Kalman filtering	Yes	No	No	No	No	Yes	No
	Group verification	Yes	No	No	No	No	Yes	Yes
	Data fusion	Yes	No	Yes	No	No	Yes	Yes
	Distance bouncing	Yes	No	No	No	No	Yes	No
	Traffic modelling	Yes	No	No	No	No	Yes	No
Secure Broadcast Authentication	PKI	Yes	No	No	Yes	Yes	Yes	Yes
	Message authentic-cation code	Yes	No	No	No	No	Yes	Yes
	Fingerprinting	Yes	No	Yes	No	No	Yes	Yes

Other methods, such as distance bounding and message authentication codes, solely address a single security need and are unacceptable as stand-alone security solutions. When considering the two tables as a whole, it should be noted that none of the proposed secure location verification mitigation strategies currently addresses eavesdropping and aircraft reconnaissance. Location verification mitigation strategies, when compared to secure broadcast authentication, are more orientated towards verifying the legitimacy of an aircraft's claimed position. In other words, it ensures location integrity. Secure Broadcast Authentication, on the other hand, are focused on authentication.

Table 2: Comparison of ADS-B countermeasures with security services

	Method	Authentication	Confidentiality	Data Integrity	Location Integrity	Availability
Secure location verification	Multi-point positioning	No	No	Yes	Yes	Yes
	Kalman filtering	Yes	No	Yes	Yes	No
	Group verification	No	No	No	No	No
	Data fusion	Yes	No	No	Yes	Yes
	Distance bouncing	No	No	No	Yes	No
	Traffic modelling	No	No	No	Yes	No
Secure Broadcast	PKI	Yes	Yes	Yes	Yes	No
	Message authentication code	Yes	No	No	No	No
	Fingerprinting	Yes	No	No	No	Yes

Maintaining confidentiality remains a challenge due to the lack of mitigation strategies that address eavesdropping and aircraft reconnaissance. Maintaining confidentiality is not as simple as one might expect. According to the US Federal Aviation Administration (FAA), operational requirements make unencrypted data links vital to ensure compatibility and global interoperability (Khandker et al, 2022b). Operational requirements for that reason make unencrypted messages and the lack of authentication one of ADS-B's primary flaws. The lack of fundamental security safeguards makes tampering with ADS-B signals easy, impacting the confidentiality, integrity, and availability of the transmitted aircraft data (Wu et al, 2020). Furthermore, only multi-point

positioning, data fusion and fingerprinting address Denial of Service attacks. As such, these are the only solutions addressing the availability of ADS-B messages and services.

Based on this analysis, it is clear that no holistic solution exists. This paper addresses this research gap by proposing a hybrid solution. Message authentication code, which is a crypto scheme more practical than PKI, should be integrated with underlying cross-verification techniques used for secure location verification. The hybrid solution would consequently address most of the ADS-B threats and thereby security services, rendering the solution acceptable as a stand-alone security solution. Furthermore, such an approach should satisfy the FAA's operational requirements. Although confidentiality is still not fully covered, it will be to a larger extent as threat actors might still intercept transmission and extract certain information. However, threat actors will not be able to tell from which aeroplane the transmission originated. The only drawback is that this solution has an underlying requirement to overcome secret key distribution challenges. That is, an aeroplane should be provided with a secret key when the specific aeroplane desires to access the air traffic control system and ADS-B network. It seems feasible since AA-MAC only needs one key to identify a sender uniquely.

5. Conclusion

The primary contribution of this paper is to not only classify the threats according to the OSI model but also to critically review how well the potential mitigation addresses the identified threats. This is accomplished by analysing both the security concerns and solutions related to ADS-B. ADS-B system attacks and mitigations have been extensively covered in published research studies. Based on the literature, ADS-B is vulnerable to cyber-attacks in multiple ways. Most of these attacks try to exploit the 1090 MHz communication channel. As such, most security solutions also relate to the 1090 MHz communication channel. Many of the proposed solutions address either a single or a small number of security concerns at a time. Furthermore, many of the solutions do not consider the limitations of the ADS-B architecture, the dense traffic on the 1090 MHz channel, or compatibility with current communications hardware and software. Consequently, the suggested countermeasures are often impractical. According to Wu et al (2020), many of the proposed solutions are only in an experimental phase. Manesh and Kaabouch (2017) continue by stating that the recommendations also call for considerable improvements. This paper addresses this research gap by proposing an alternative solution. Future research will focus on the development of the hybrid solution presented in Section 4.

References

Bishop, G. & Welch, G. (2001) *An introduction to the kalman filter*. SIGGRAPH, Course, 8(27599-23175):41.

Blåberg, A., Lindahl, G., Gurtov, A. & Josefsson, B. (2020) *Simulating ADS-B attacks in air traffic management*. 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). IEEE. pp. 1-10.

Giannatto Jr, C.J. 2015. *Challenges of implementing automatic dependent surveillance broadcast in the nextgen air traffic management system*.

Hasin, F., Munia, T.H., Zumu, N.N. & Taher, K.A. (2021) *ADS-B based air traffic management system using ethereum blockchain technology*. 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD). IEEE. pp. 346-350.

Khandker, S., Turtiainen, H., Costin, A. & Hämäläinen, T. (2021) *Cybersecurity attacks on software logic and error handling within ADS-B implementations: Systematic testing of resilience and countermeasures*. IEEE Transactions on Aerospace and Electronic Systems, 58(4):2702-2719.

Khandker, S., Turtiainen, H., Costin, A. & Hämäläinen, T. (2022a) *On the (in) security of 1090ES and UAT978 mobile cockpit information systems–an attacker perspective on the availability of ADS-B safety-and mission-critical systems*. IEEE Access, 10:37718-37730.

Khandker, S., Turtiainen, H., Costin, A. & Hämäläinen, T. (2022b) *Cybersecurity attacks on software logic and error handling within ais implementations: A systematic testing of resilience.* IEEE Access, 10:29493-29505.

Li, W. & Kamal, P. (2011) *Integrated aviation security for defense-in-depth of next generation air transportation system*. 2011 IEEE International Conference on Technologies for Homeland Security (HST). IEEE. pp. 136-142.

Manesh, M.R. & Kaabouch, N. (2017) *Analysis of vulnerabilities, attacks, countermeasures and overall risk of the Automatic Dependent Surveillance-Broadcast (ADS-B) system*. International Journal of Critical Infrastructure Protection, 19:16-31.

McCallie, D., Butts, J. & Mills, R. (2011) *Security analysis of the ADS-B implementation in the next generation air transportation system*. International Journal of Critical Infrastructure Protection, 4(2):78-87.

Mirzaei, K.F., De Carvalho, B.P. & Pschorn, P. (2019*) Security of ADS-B: Attack scenarios*. EasyChair, Tech. Rep,

Naganawa, J. & Miyazaki, H. 2018. *A method for accurate ads-b signal strength measurement under co-channel interference*. 2018 Asia-Pacific Microwave Conference (APMC). IEEE. pp. 354-356.

Schäfer, M., Lenders, V. & Martinovic, I. (2013) *Experimental analysis of attacks on next generation air traffic communication*. Applied Cryptography and Network Security: 11th International Conference, ACNS 2013, Banff, AB, Canada, June 25-28, 2013. Proceedings 11. Springer. pp. 253-271.

Viveros, C.A.P. (2016) Analysis of the cyber attacks against ADS-B perspective of aviation experts.

Wu, Z., Shang, T. & Guo, A. (2020) *Security issues in automatic dependent surveillance-broadcast (ADS-B): A survey*. IEEE Access, 8:122147-122167.

Ying, X., Mazer, J., Bernieri, G., Conti, M., Bushnell, L. & Poovendran, R. (2019) *Detecting ADS-B spoofing attacks using deep neural networks*. 2019 IEEE conference on Communications and Network Security (CNS). IEEE. pp. 187-195.

Zeng, K., Govindan, K. & Mohapatra, P. (2010) *Non-cryptographic authentication and identification in wireless networks [security and privacy in emerging wireless networks]*. IEEE Wireless Communications, 17(5):56-62.

Identifying the Scope of Cybersecurity Research Conducted in the Maritime Industry: 2003 - 2023

Tshepo Mawer[1,2], Sune von Solms[1,2] and Johan Meyer[1]
[1]Department of Electrical and Electronic Engineering Science, University of Johannesburg, South Africa
[2]South African International Maritime Institute, Nelson Mandela University, Gqeberha, South Africa

tmawer@uj.ac.za
svonsolms@uj.ac.za
johanm@uj.ac.za

Abstract: The maritime industry plays a pivotal role in the modern economy, global trade, safety and transportation globally. Due to the rapid advancement and utilisation of technology in the field, the maritime industry is increasingly vulnerable to cybersecurity threats. The last decade saw many instances of infrastructure having been exploited or attacked, such as the maritime vessels themselves, the port infrastructure and the supply chain. Each component constituting the maritime industry requires unique system critical operations to ensure cybersecurity, however this paper focuses on certain aspects of the maritime industry only. As cybersecurity in the maritime industry is a growing field, this study determines the scope of cybersecurity research on maritime infrastructure which has been published to establish a baseline understanding of the current field. The paper presents the results from a systematic literature review conducted to assess the scope of current cybersecurity work published focusing on the maritime industry. The result from the study clearly shows the increased attention to cybersecurity in the maritime industry, which can be seen from the increased number of publications which showed a sharp increase after 2014.

Keywords: Cybersecurity, Maritime, Information security

1. Introduction

The history of seafaring and boat use dates to ancient times, serving various purposes like civilization mobility, exploration, trade, warfare, and fishing. Maritime sailing originated around the 3rd century BC, proving faster and more cost-effective for product transportation than land (Shipfinex, 2023). Container ships now move 90-95% of global manufactured goods, with the 2019 world shipping trade valued over $14 trillion (Di Fonzo & Paris, 2018). Navies and coast guards ensure safety in territorial waters, trade routes, and the environment, making the maritime industry a crucial global transport mode (Shipfinex, 2023).

Maritime activities significantly impact the world economy. The Suez Canal blockage in 2021 disrupted the global supply chain (Shipfinex, 2023). The maritime sector, undergoing modernization, faces growing cybersecurity threats with far-reaching consequences. It includes vessels, port infrastructure, and the supply chain, necessitating cybersecurity to ensure safety. Information and Communication Technology (ICT) is increasingly adopted for cost-effectiveness, safety, and sustainability, making cybersecurity a priority (Moan, 2022). Properly securing and managing ICT systems aboard ships, in ports, and throughout the supply chain is essential (Moan, 2022).

Cases of insecure systems or insecure operation have led to cybersecurity attacks, such as:

- A British tanker, Stella Impero, in the Strait of Hormuz took a sharp turn off course. It is then seized by Iranian forces in their waters. The ship's GPS location was spoofed. (July 2019).
- A cyber-attack against Transnet, who operates two major ports and most railways in South Africa, disrupted container operations at the ports of Cape Town and Durban. Operation capacity was reduced to 10% of normal operations, forcing workers to process cargo manually (July 2021) (Heiberg T., 2021).

With the increase of above-mentioned types of cyber incidents, the focus on cybersecurity in the maritime industry grows, and the question that emerges is whether the status of maritime cybersecurity in any manner reflects the production of scientific publications in this field? This study aims to analyse the scope of scientific publications in the field of cybersecurity in the maritime environment. This study provides an indication of the standing of cybersecurity research in the maritime environment that can serve as a baseline towards understanding the extent and scope of research conducted in the field. The paper is structured as follows: Section 1 includes the introduction where Section 2 provides a brief overview on the history of cybersecurity

and the maritime landscape. Section 3 provides a discussion on the methodology followed in this study with the results following in Section 4. Section 5 concludes this paper.

2. History of Cybersecurity and the Maritime Industry

The maritime industry continuously adapts to evolving technology, with ICT systems and advanced technology profoundly shaping a more interconnected and technology-dependent sector (Apud, 2016; Lind-Olsen, 2019). It involves advanced navigation systems using GPS, radar, sonar, computerized maps, control systems for electromechanical systems, radio and satellite-based communications, and Automatic Identification Systems (AIS). Modern ships, ports, and supply chains generate extensive data shared among various stakeholders (Apud, 2016; Lind-Olsen, 2019; Akpan, 2022; Alcaide, 2020).

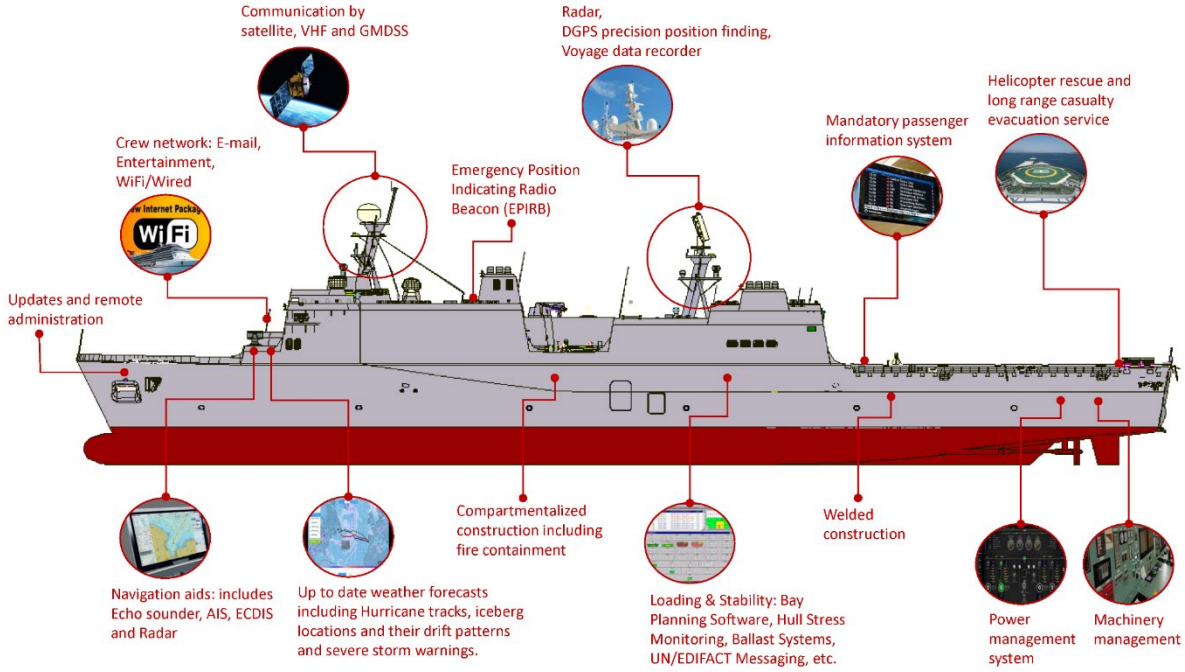


Figure 1: Automation systems for modern and autonomous ships (Akpan et al, 2022; CruisMapper, 2022)

The global rise in e-commerce and global shopping creates new opportunities for the maritime supply chain, not just for shipping and ports (UNCTAG, 2021). In 2020, the global commercial shipping fleet grew by 3%, totalling 99,800 ships with a gross tonnage of 100 or more. By January 2021, the combined capacity reached 2.13 billion dead weight tons. In 2020, ship deliveries fell by 12% due in part to lockdown-induced labour shortages, disrupting maritime-industrial operations. Most of the new ships were bulk carriers, followed by oil tankers and container vessels. To cope with limited vessel supply, owners and operators acquired second-hand ships with a mix of legacy and modern systems (UNCTAG, 2021).

Data generation and sharing as well as the utilisation of technology is not something new and plays a major role in many industries globally. However, the maritime industry and the related operations are unique because the maritime environment (Mouton, 2023):

- Spans a vast operational environment and limited connectivity.
- Includes the integration of operational technology (OT) and information technology (IT) systems.
- Includes a multitude of vessels and ports which uses legacy systems.
- Has a lack of standardisation and regulatory challenges.
- Suffers from supply chain vulnerabilities and insider threats.
- Suffers from limited awareness and training.

This unique environment raises cybersecurity issues for both modern and legacy systems. As more devices and systems are introduced into the broader network, the windows for cyber-attacks increase (Ignacio, 2021). This heightened connectivity renders the system more vulnerable to diverse cyber threats. In 2013, hackers took control of an oil rig's stabilizing system in the Gulf of Mexico, leading to a 19-day shutdown to rectify the dangerous tilt caused by the attack. In 2016, 280 Korean vessels experienced GPS jamming, resulting in navigation system shutdowns and receipt of false information (Chang, et al., 2019).

Cyber threats in the maritime industry are further compounded by the awareness of those involved in this space, emphasizing the need for additional research to address these vulnerabilities (Afenyo & Ceasar, 2023). While many companies within the maritime sector are working to enhance their cybersecurity measures, both the industry and academia recognize gaps in research, policy, standards, and personnel training (Afenyo & Ceasar, 2023). The maritime industry should develop cybersecurity protocols, enhance education and awareness, and promote further research to build on existing knowledge (Afenyo & Ceasar, 2023) (Chang, et al, 2019).

According to BIMCO guidelines (BIMCO et al., 2021; UNCTAG, 2021):

- Recognize cybersecurity threats, including internal and external risks.
- Identify asset weaknesses by listing onboard systems with communication links, ensuring all parties understand threats and protective measures.
- Evaluate risk exposure, vulnerability, and potential exploitation.
- Develop protective and detection strategies to reduce vulnerability impact.
- Create response plans with contingency measures for cyber risks.
- Efficiently respond and recover from incidents, assess response plan effectiveness, and update for changing threats (BIMCO et al., 2021).

As the focus on cybersecurity in the maritime industry is increasing, this paper tries to determine if the focus of cybersecurity in the maritime industry reflects in the production of scientific publications in this field? In order to determine the scientific publications published on cybersecurity in the maritime industry, a systematic literature review was conducted to determine the trends in research publications in the field.

3. Methodology and Information Sources

A systematic literature review (SLR) is a comprehensive and methodological approach to identify and collate publications and research that have already been conducted, tailored towards answering a specific research question or hypothesis (Xiao & Watson, 2019). The decisions and procedures for including or excluding certain publications are to be well defined and/or justified, to ensure that the review is repeatable (Mohamed Shaffril et al., 2020). Through the lens of a SLR one can discover the depth and breadth of the existing body of knowledge pertaining to a topic or research question, while identifying gaps that require further exploration (Xiao & Watson, 2019). By rephrasing the aim of this paper, the authors determined the research objective necessary for the systematic literature review – *To determine the scope of research published within the field of cybersecurity and the maritime industry.*

The PICO (Population, Intervention, Comparators, Outcome) methodology for conducting a systematic literature review (SLR) was used. This methodology entails defining the research question around the four parameters (from which PICO is derived) and utilising that framework to develop the search terms or combination of keywords that will be used in the database search (Purssell & McCrae, 2020). The methodology assists in narrowing the focus of the SLR to more of the relevant documentation, rather than focusing the research net too wide. A variation of this model was chosen, specifically the Population, Exposure and Outcome (PEO) method due to its focus on the exposure of the research rather than its intervention, as the aim of this paper was to establish what research had already been conducted (Purssell & McCrae, 2020). For this paper the PEO model looks as follows:

Population: This parameter represents the sample space that is being explored, which for this paper would be the *maritime industry.*

Exposure: This parameter represents the research within the maritime field that are cybersecurity related that have already been published, which for this paper would be the *cybersecurity.*

Outcome: This parameter represents the outcome which this study necessitates, which is to develop an understanding of the current *scope of research* that has taken place (as of this paper's publication) into *cybersecurity within the maritime context.*

In the SLR, the considered publications were restricted to journals, articles, and peer-reviewed conferences meeting specific criteria, including international reputation and English language. Two databases were chosen to ensure insightful results: IEEE Xplore, a digital library by the IEEE, covering engineering, electronics, and

computer science (IEEE Xplore, 2023); and Scopus, a multidisciplinary digital library with peer-reviewed literature, including journals, books, and conference proceedings (Elsevier, 2023). Scopus is the largest citation and abstract database with smart research analysis tools. IEEE Xplore and Scopus were selected due to their integrated tools for capturing and analysing published research (Elsevier, 2023).

The SLR process followed can be described by the following steps:

Step 1. The period of data considered for the SLR was not restricted.

Step 2. Research publications in English, peer-reviewed and of type journals, articles and conference papers were considered.

Step 3. Table 1, below gives the key words selected for the SLR.

Table 1: Table of search terms and their synonyms

Keyword	**Synonyms**				
cybersecurity	"cyber security"	"cyber-security"	"network security"	"information security"	"security of data"
maritime	marine				

For each database, two searches were conducted, shown below. The use of Boolean operators was vital in ensuring a controlled search across both databases, as well as the use of quotation marks in order to restrict the search to the relevant keywords/search terms.

String 1: (cybersecurity OR "cyber security" OR cyber-security OR "network security" OR "information security" OR "security of data") AND (maritime)

String 2: (cybersecurity OR "cyber security" OR cyber-security OR "network security" OR "information security" OR "security of data") AND (marine)

The first search included maritime as a keyword, which refers to anything related to shipping, navigation, or commerce of the sea (such as maritime law and maritime trade) (Youth Maritime Collaborative, 2023). Within the context of research, the term maritime is often linked to the design, innovations and technologies meant to enable better exploration, transportation and human endeavour upon the oceans and seas (Hildebrand & Schröder-Hinrichs, 2014). The second search replaced maritime with marine, which refers to anything related to the ocean or sea, such as marine biology or life (Hildebrand & Schröder-Hinrichs, 2014). Within the context of research the term marine often refers to the understanding and study of ecologies and ecosystems, human impact on those ecosystems and avenues to move forward with to enable sustainable use and growth NOAA, 2022).

Step 1. Publication search: The title, abstract and keywords of each publication in the search database were interrogated for the presence of the search terms. When the search terms were found, the following meta information from the publication was captured in the results database which included:

- Title of paper
- Year of publication
- Authors and affiliations
- Author and Index Keywords
- Name of conference proceedings or journal
- Abstract
- Digital Object Identifier (DOI)

Step 2. Exclusions: The following exclusion criteria were applied to the initial search results. Publications with primary focus on the terms indicated in Table 2, were excluded for the reasons stated:

Table 2: Exclusion terms

Exclusion term	**Reason for exclusion**
Ports and port infrastructure	Not in scope, focused on maritime vessels
Naval and naval vessels	Focused on commercial/civilian vessels, not military
Oil rigs/freight/container	Cybersecurity in the offshore exploration domain was not the focus of this study

Underwater communication networks/ submarine cables	Underwater communication systems utilised in the communication industry were not the focus of this study
Exclusion term	**Reason for exclusion**
Energy generation	Offshore energy generation platforms and systems were not considered
Inland rivers	Excluded research on inland shipping
Environmental monitoring systems **Internet of Underwater Things** **Underwater sensor systems**	Systems and sensors utilised for the collection of data regarding the marine environment were excluded
Marine farming	Publications with a focus on marine farming were excluded as the focus of this study was the maritime vessel
Laws, strategies and policies	Publications with primary focus on maritime policy or maritime law were excluded from this study

Step 3. Duplication Removal: Duplicate publications from the database search results were eliminated using two methods. Initially, the unique DOI was compared to remove duplicates. The second pass involved cross-referencing document titles and author names to further eliminate any remaining duplicates.

Note that this search does not include corporate or military research into cybersecurity in the maritime industry as this study considers the contribution of academic research from academic institutions. This search is not an exhaustive study as not all academic databases were considered in this study. However, the search was considered to be of sufficient size in order to provide trends in maritime cybersecurity publications.

4. Results

The data collection methods discussed in Section 3 were applied using the IEEE Xplore and Scopus databases. After an initial search the total amount of publications before scope and duplicate exclusions totalled 1336 publications. A total number of 487 publications spanning from the year 2003 till 2023 were obtained after out of scope and duplicate exclusions were removed. To obtain an overview of the scope of research conducted on cybersecurity in the maritime industry a number of data depictions were generated, each providing different insight into the domain of cybersecurity research in the maritime industry.

4.1 Number of Publications per Year

The number of publications per year in the maritime industry focussed on cybersecurity is an indicator of the interest given and research effort allocated to this field. The number of publications generated per year is shown in Figure 2.

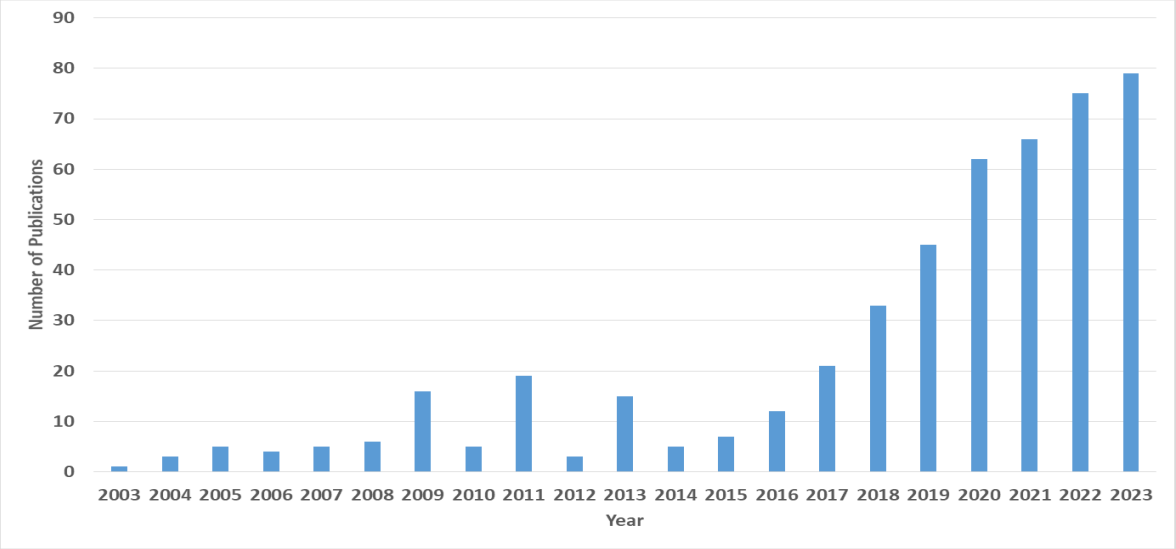


Figure 2: Number of publications per year over the entire study period from 2003 to 2023

The number of publications in the search field as function of year published is shown in Figure 2. The results showed that from the year 2003 until 2014 an average of about 7 conferences papers and/or articles were

published. From 2003 until 2008 the largest number of publications was a total of 6 publications in 2008. From the year 2008 until 2014 there were alternating years of similar amounts of publications and spikes in the amount of research published, with 2011 holding the largest number for that period of 19 publications. However, if one looks at the results from 2015 until 2023 an obvious trend is established, as there was a significant increase in the number of publications over that period of time. Starting with 7 publications in 2015 the number of publications increases to 79 publications so far for 2023. These results indicate a significant increasing trend in cybersecurity research in the domain of the maritime industry. An exponential increase of nearly 1029% is observed over the period from 2015 to 2023. The aim of this research was not to come up with a prediction of where these numbers might go in future therefore, a regression curve was not fitted to the data presented in Figure 2.

It is insightful to compare the growth in publications in cybersecurity in the maritime industry with the growth in the industry itself. The international maritime trade consisting of tanker, bulk and other dry cargo, has grown from 5 984 million tons loaded per year to a maximum of 11 071 million tons loaded in 2019 (United Nations, 2021). Although the maritime tons loaded more than doubled, the increase in cybersecurity around the maritime industry has increased more than tenfold showing how the interest in cybersecurity around the maritime industry has intensified as shown in Figure 2.

4.2 Number of Publications per Country

An analysis of the country affiliation of the authors from the publications provide insight into the countries conducting or collaborating in research related to cybersecurity in the maritime industry. Figure 3 shows the number of publications obtained from the SLR ranked by country affiliation of the authors.

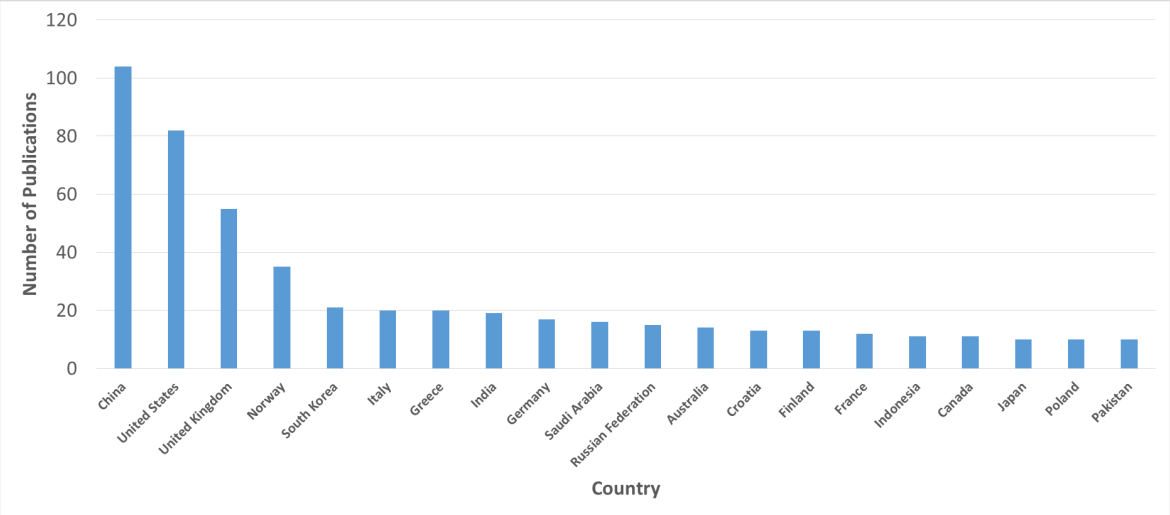


Figure 3: Number of publications per country over the entire study period.

The percentage share per region in world maritime trade (United Nations, 2021), may be indicative of the countries producing research in maritime cybersecurity. Asia with the highest percentage share (54% trade share) has also produced the highest number of publications namely 104 in total from China affiliated researchers with the Americas (18% trade share) second with 82 publications followed by European countries (15 % trade share) such as the United Kingdom and Norway producing 55 and 35 publications respectively. It is interesting to note that the combined number of research publications from China, United States and United Kingdom and Norway forms 41.69% of the total amount of publications produced globally.

Indicating the research publications numbers as generated per country on a global country heat map, provides a graphical indication of geo-located research in maritime cybersecurity. Figure 4 is a heat map of the number of research publications obtained by the SLR, geo-located on the global map.

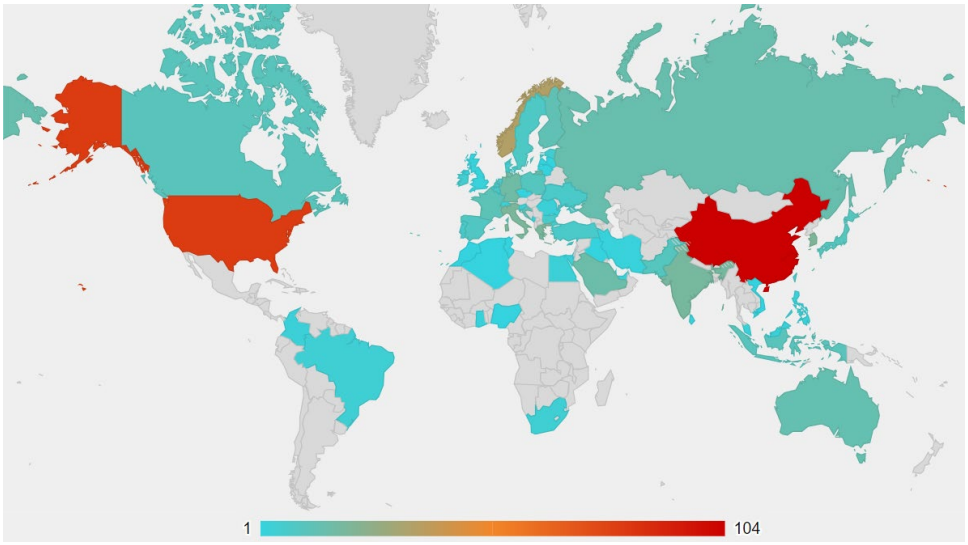


Figure 4: Number of research publications per country.

4.3 Cybersecurity Topics of Interest

Insight into the underlying topics of interest which was researched can be obtained by looking at the key words used to describe the research publications. Table 3 gives the frequency histogram ranked by occurrence, of the keywords used to describe the topics of interest.

Table 3: List of key word used to describe the research topics

Keyword	Frequency	Keyword	Frequency	Keyword	Frequency
Network security	203	Internet of Things	45	Marine transportation	27
Cybersecurity	152	Waterway transportation	41	Marine communication	23
Ships	128	Crime	36	Cryptography	22
Security of data	115	Risk management	34	Maritime industry	21
Risk assessment	81	Automatic Identification System	30	Maritime security	21
Security	64	Computer crime	28	Authentication	20
Cyber-attacks	50	Automation	27	Monitoring	20



Figure 5: Word cloud map of keywords over the full 21 years study period

It is no surprise that the top ranking three keywords are Network security (203), Cyber Security (152) and Ships (128). However, more insight into the underlying research field can be obtained by looking at the lesser ranked keywords. A word cloud with the word size as an indication of frequency of occurrence is shown in Figure 5. Using this means of representation is useful since it provides insight into the overall field of research in one glance. The lesser used topics of research are more easily represented in this form, and can be more easily compared to the more popular research topics.

4.4 Keywords per Year

An analysis of keyword usage per year gives an indication of the evolution of the research topics over the research period. Figure 6 shows how the emphasis on particular keywords used evolved over the study period. As the number of publications increased from the year 2015 so did the growth in the usage of the primary keywords.

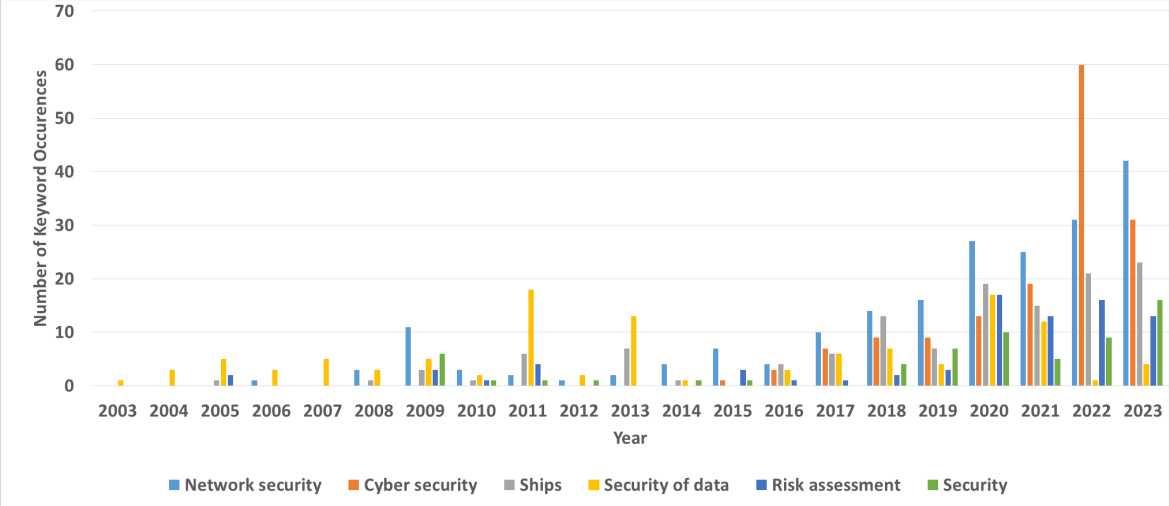


Figure 6: Number of keywords per year over the entire study period. For better visualisation the amount of keywords displayed is restricted to the top five keywords, as determined by the number of times those keywords appeared over the entire study period.

4.5 Top Keywords per Focus Area

An analysis of keyword usage per year gives an indication of the evolution of research topics over the research period. Figure 7 shows how the emphasis on particular keywords used evolved over the study period.

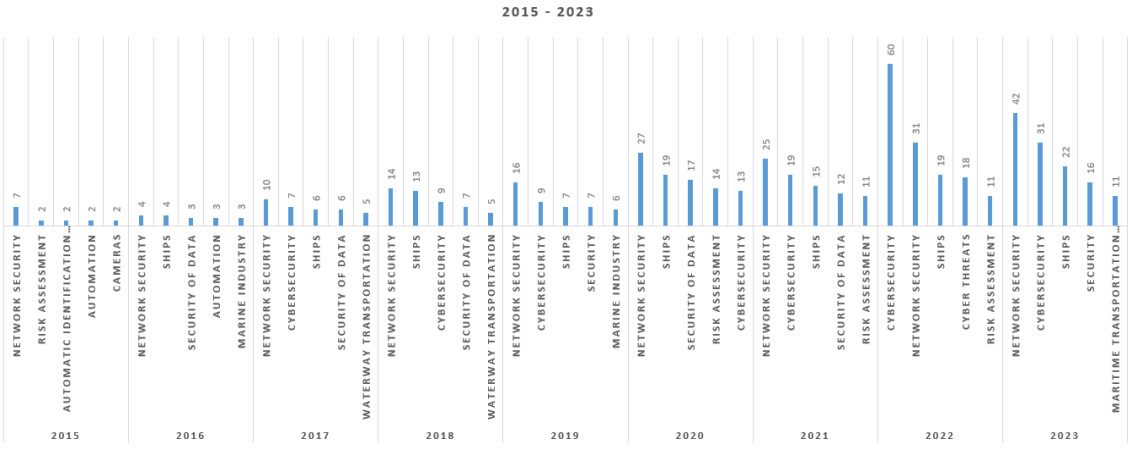


Figure 7: Growth in keyword usage over the period 2003 to 2023

From the year 2003 to 2015 there was a focus on the security of data and network security, through the appearance of keywords such as “Security of data”, “Computer networks”, “Risk assessment” and “Computer crime”. Other keywords such as “Global positioning system”, “Coastal zones”, “Marine engineering”, “Waterway transportation”, “Maritime security”, “Supply chains” and “Automation” can be found in the most frequently used search terms across those years as well. This indicates that cybersecurity and its maritime applications

were still in an earlier stage of development, as not only was there a smaller amount of research being conducted within these focuses, the content that did appear was primarily on specific systems within the maritime industry (such as those used in navigation or automation) as well as general risk assessment.

From 2016 to 2023 the top three keywords per year were always “Network security”, “Cyber security” and “Ships”, although the frequency with which they were used and the order in which they appear changed with time. This time period coincides with the surge of growth in research over the entire study period, and now it is clear that the focus area of many of those studies was specifically the protection of digital data and systems regarding ships. Other keywords that appeared frequently during this timeframe include “Automation”, “Waterway transportation”, and “Risk assessment”. This indicates that the need for research within cybersecurity and water transportation is still present, just the focus has shifted with time to be more general.

A similar trend can also be observed through the number of keywords used across each year of publications. Over the entire 20 year period for this study, the keywords used were tallied (which can be visually represented in Figure 5, as a word cloud map). “Network security” was the search term to appear the most, totalling 203 times. “Cybersecurity”, “Ships” and “Security of Data” showed up the most frequently after that, appearing 152, 128 and 115 times respectively. By separating these results by year it became clear that there was a significant increase in the number of these keywords from 2015 to 2023, as seen in Figure 7. These results follow a similar trend to that of the number of publications across that time period. Figure 7 also revealed that not every keyword was used over the whole 2003 - 2023 time period, with a large proportion of the keywords only being used frequently from about 2016.

5. Discussions and Conclusions

When reflecting on the history of cybersecurity in the maritime landscape it is clear that there has been an increase in the cyber-threats to the maritime industry over the last 21 years. The technology and systems being used in the maritime industry have been rapidly developing and evolving, however the cybersecurity systems in place have not developed as rapidly, or the integration of more modern systems with the existing legacy systems have led to greater opportunities for cybersecurity threats to exploit.

Through the lens of a systematic literature review this paper was able to gain insights into the scientific publications that have been developed over this period of rapid change. The number of publications per year over the period of study indicate a significant increase in research publications from about 2014, as well as an increase in the keywords associated with the maritime industry and cybersecurity in general. The most frequently appearing keywords being “Network security”, “Cyber security” and “Ships”, all of which saw a significant rise in used from 2015. The SLR also indicted that a vast majority of the research that had been published was from the United States, China, the United Kingdom and Norway. There was also a noticeable lack in published research from African countries as a whole.

In conclusion, when comparing the history of cybersecurity within the maritime industry to the insights and patterns found from the results of the systematic literature review it is apparent that there has been an increase in the attention paid to cybersecurity within the maritime industry, especially from 2014. This attention is reflected in the sharp increase in the number of scientific publications that have been produced in more recent years, as well as the significant increase in the use of keywords directly linked to cybersecurity and the maritime industry.

References

Alcaide, J.I.; Llave, R.G. (2020) “Critical infrastructures cybersecurity and the maritime sector”. Transp. Res. Procedia 2020, 45, 547–554.

Akpan, F., Bendiab G., Shiaeles, S., Karamperidis, S., Michaloliakos, M., (2022) “Cybersecurity Challenges in the Maritime Sector”, Network 2022, 2(1), 123-138

Apud, Julius Patrick, (2016) “Information Technology Applications in the Maritime Industry”, The Maritime Review, [online], https://maritimereview.ph/information-technology-applications-in-the-maritime-industry/

Chang, C-H, Wenming, S, Wei, Z, Changki, P and Kontovas, CA (2019) “Evaluating cybersecurity risks in the maritime industry: a literature review” International Association of Maritime Universities (IAMU), [online], https://researchonline.ljmu.ac.uk/id/eprint/11929

CruisMapper. Cruise Ship Safety. Available online: https://www.cruisemapper.com/wiki/751-cruise-ship-safety (accessed on 3 February 2022).

Di Fonzo, T. & Paris, LLC. (2018) “How a Steel Box Changed the World: A Brief History of Shipping”, WSJ Video, [online], https://www.wsj.com/video/series/a-brief-history-of/how-a-steel-box-changed-the-world-a-brief-history-of-shipping/CF460889-9984-483E-AF44-324330B89ECA

Elsevier (2023) “About | Elsevier Scopus Blog”, [online], https://blog.scopus.com/about#:~:text=Scopus%20is%20a%20source%2Dneutral,promote%20ideas%2C%20people%20and%20institutions

Heiberg T. (2021) “Transnet restores operations at ports after cyber attack”, CNBC Africa, [online], https://www.cnbcafrica.com/2021/transnet-restores-operations-at-ports-after-cyber-attack/

Hildebrand, L.P., Schröder-Hinrichs, JU. (2014) “Maritime and marine: synonyms, solitudes or schizophrenia?” WMU J Marit Affairs 13, 173–176 (2014). https://doi.org/10.1007/s13437-014-0072-y

IEEE Xplore (2023) “About IEEE Xplore”, [online], https://ieeexplore.ieee.org/Xplorehelp/overview-of-ieee-xplore/about-ieee-xplore

Ignacio de la Peña Zarzuelo, (2021) “Cybersecurity in ports and maritime industry: Reasons for raising awareness on this issue”, Transport Policy, ISSN 0967-070X, https://doi.org/10.1016/j.tranpol.2020.10.001

Lind-Olsen, Morten, (2019) “ICT solutions bring ship and shore closer”, Dualog, [online], https://www.dualog.com/blog/ict-solutions-bring-ship-and-shore-closer

Mawuli Afenyo, Livingstone D. Caesar, (2023) “Maritime cybersecurity threats: Gaps and directions for future research”, Ocean & Coastal Management, ISSN 0964-5691, https://doi.org/10.1016/j.ocecoaman.2023.106493

Moan, S. (2022) “The value of ICT in the maritime industry”, Dualog, [online], https://www.dualog.com/blog/the-value-of-ict-in-the-maritime-industry

Mohamed Shaffril, H. A., Samsuddin, S. F., & Abu Samah, A. (2020) “The ABC of systematic literature review: the basic methodological guidance for beginners”, Quality and Quantity, 0123456789. https://doi.org/10.1007/s11135-020-01059-6

Mouton, F. (2023) “Charting a secure voyage: Navigating social engineering in maritime”, NORMA Cyber Conference.

National Oceanic and Atmospheric Administration (NOAA), U.S Department of Commerce, (2022) “Marine Scientific Research”, NOAA, https://www.noaa.gov/marine-scientific-research

Paré, Guy, Marie-Claude Trudel, Mirou Jaana, and Spyros Kitsiou. (2015) “Synthesizing Information Systems Knowledge: A Typology of Literature Reviews.” Information & Management, 52:183–99. https://doi.org/10.1016/j.im.2014.08.008

Purssell, E. and McCrae, N. (2020) “How to Perform a Systematic Literature Review, How to Perform a Systematic Literature Review”, ISBN 978-3-030-49671-5. https://doi.org/10.1007/978-3-030-49672-2

Shipfinex, (2023) “Emphasis Of Maritime Industry On The World Economy”, LinkedIn [online] https://www.linkedin.com/pulse/emphasis-maritime-industry-world-economy-shipfinex/

UNCTAD (2021), “Review of Maritime transport”, https://unctad.org/system/files/official-document/rmt2021_en_0.pdf

United Nations. (2021) “Review of Maritime Transport 2021” ISBN: 978-92-1-113026-3 https://unctad.org/system/files/official-document/rmt2021_en_0.pdf

Xiao, Y., & Watson, M. (2019) “Guidance on Conducting a Systematic Literature Review. Journal of Planning Education and Research”, 39(1), 93-112. https://doi.org/10.1177/0739456X17723971

Youth Maritime Collaborative (2023) “What is Maritime?” [online], https://youthmaritimecollaborative.org/what-is-maritime

National Critical Information Infrastructure Protection Through Cybersecurity: A National Government Perspective

Thuli Mkhwanazi and Lynn Futcher
Nelson Mandela University, Summerstrand, South Africa
Council of Scientific and Industrial Research (CSIR)

S216316987@mandela.ac.za
lynn.futcher@mandela.ac.za

Abstract: The South African national government is forging ahead with digitalisation plans to enhance socioeconomic growth in the country. However, digitalisation is accompanied by detrimental cybersecurity risks that may potentially exacerbate the vulnerability of the National Critical Information Infrastructure (NCII) of South Africa to cyber-attacks. Therefore, the envisaged digitalisation benefits in South Africa may be offset by the increased cybersecurity risk on the NCII of the country.

Through a theoretical literature review, this study aims to investigate digitalisation and identify the cybersecurity risks it poses to the National Critical Information Infrastructure of South Africa, from a national government policy perspective. A gap has been identified in Information and Communications Technology for Development (ICT4D) research studies that researchers tend to focus on the implementation of ICTs while neglecting the policy aspect that is meant to direct and control the implementation of ICTs. Therefore, this study bridges this gap by approaching digitalisation and cybersecurity from a national government policy perspective. The output of this study is a National Critical Information Infrastructure Protection (through Cybersecurity) Conceptual Framework (NCIIP-CF) that is applicable to all spheres of government (local, provincial or national) as policy makers. The NCIIP-CF demonstrates an approach that embeds cybersecurity in the digitalisation process for the national government of South Africa, thus enabling an NCII that is resilient to cybersecurity risks.

Keywords: Digitalisation, Fourth Industrial Revolution, Cybersecurity, South African Government, National Cybersecurity, National Critical Information Infrastructure

1. Introduction

Digitalisation, also known as the "Fourth Industrial Revolution" or "Industry 4.0," is a global megatrend (Department of Science and Technology, 2019; Thun et al., 2019) that has that gained attention from influential international countries and organisations, like the World Economic Forum (WEF) (Cunningham, 2018), Group of 20 (G20 Osaka Leaders, 2020), and Brazil, Russia, India, China and South Africa bloc (BRICS Heads of State and Government, 2018). It aims to amplify socioeconomic productivity through the diffusion of emerging technologies (Bican and Brem, 2020). The COVID-19 pandemic accelerated the global transition to digitalisation, leading to the South African government prioritizing digitalisation plans (Schwab, 2020; Wang et al., 2019).

However, digitalisation presents cybersecurity risks, particularly on the National Critical Information Infrastructure (NCII) (Abdullah et al., 2019; Cunningham, 2018; Laufs et al., 2020; Victoria, 2019). National governments are responsible for protecting the NCII through cybersecurity measures. However, the South African government has not fulfilled this mandate successfully, with several cyber-attacks causing significant economic loss (Griffiths, 2017; Sutherland, 2017; Van Niekerk, 2017).

This paper proposes an NCII Protection Conceptual Framework (NCIIP-CF) to mitigate these risks, prioritizing national cybersecurity in the digitalisation process from the beginning, which is the public policy planning phase.

This paper outlines the structure of digitalisation and its critical components (Section 2), the NCII (Section 3), NCII protection through cybersecurity (Section 4), and a conclusion (Section 5).

2. Digitalisation

Digitalisation is the process of transforming world systems (social, economic and political) into a borderless domain, through digitisation. Digitisation is the technological process of transforming physical or manual systems, services or products from a static format to an interactive and autonomous one that can be stored, processed and transmitted electronically using extensive computational power (Bican and Brem, 2020). Table 1 presents some of the emerging technologies being referred to when speaking of digitisation (Bican and Brem, 2020; Pillay et al., 2016; Victoria, 2019). Digitisation combines interdependent physical, digital and biological elements into novel products and services (Schwab, 2016).

Table 1: Digitisation Emerging Technologies (Bican and Brem, 2020; Cunningham, 2018; Mitchell, 2020; Schwab, 2016)

Physical emerging technologies	Digital emerging technologies	Biological emerging technologies
3D Printing	IoT	Genetic sequencing
Autonomous vehicles	AI	Gene edition or activation
New material	Blockchain	Synthetic biological cells
Robotics	Virtual and augmented reality	Neurotechnology
Quantum computers	5G	
Space technologies	Social media	
	Cloud technology	

Widespread digitisation will, in turn, transform the entire structure of interaction between humans and the world systems, which is a phenomenon known as digitalisation.

2.1 Digitalisation Theoretical Framework

Digitalisation is not rooted in Information and Communication Technologies (ICT) as common sense would argue, but rather in historic economic cycles (Hilbert, 2012). Conceptually, economic cycles are wave-like occurrences, where the incline represents the rising innovation which increases industrial productivity, profits and economic growth until a point of equilibrium is reached (Perez, 1985; Schumpeter, 1939). The point of equilibrium occurs when the innovation has served its purpose and ceases to be novel; thus, leading to a decline in industrial productivity, profits and economic growth, moving into a recession phase (Schumpeter, 1939). A recession phase calls for innovation and thus for another economic cycle (Perez, 1985; Schumpeter, 1939). There are various economic cycles, including 'Kondratieffs, Kitchins and Juglars' (Schumpeter, 1939, p. 176). Schumpeter deduced that, of the above-mentioned three economic cycles, the Kondratieff is an industrial revolution (IR) (Schumpeter, 1939). An IR is a group of related and interdependent innovations that, when largely adopted in industry, bring about a transformation in the broader social and economic structures of a nation (Schumpeter, 1939). Schumpeter's description of an IR fits the description of digitalisation. Hence this study regards the fourth IR (4IR) as the theoretical framework of digitalisation.

The next section discusses ICT for Development (ICT4D) as a conceptual framework for digitalisation.

2.2 Digitalisation Conceptual Framework

From a high-level overview of digitalisation, the themes of emerging digital ICT and socioeconomic development can be deduced. This is drawn from the concept of ICT4D (Ordóñez, 2015). ICT4D entails projects that are conducted by donors to establish or to improve the state of ICT systems of recipients, for socioeconomic development (Chipidza and Leidner, 2019). Depending on the locale of the ICT4D project, the donors are the initiators of the ICT4D project as they are more capable of doing so owing to the availability of resources to them (Hilbert, 2012). Typically, donors are international influencers who are superior in terms of socioeconomic development (Kunyenje, 2019). Therefore, the likes of BRICS, G20, WEF, and United Nations (UN), referred to in Section 1 represent donors taking up the role of assisting the recipients. Donors are not limited to influential countries — they could be researchers, governing authorities, private banks or any stakeholder taking up the role of assisting the recipients (Chipidza and Leidner, 2019; Hilbert, 2012). The recipients are the beneficiaries or subjects who will supposedly benefit from the ICT4D project, be it a government, public sector, community or typically, a developing country like South Africa (SA) (Chipidza and Leidner, 2019; Schia, 2018). As such, digitalisation qualifies to be viewed through the lens of ICT4D in developing countries since it comprises the harnessing of emerging digital ICT for socioeconomic development, and it is promoted by donors.

Figure 1 presents the ICT4D-CF that was originally established by the United Nation's Economic Commission for Latin America and the Caribbean (UN-ECLAC) and modified by Hilbert in 2012 (Hilbert, 2012).

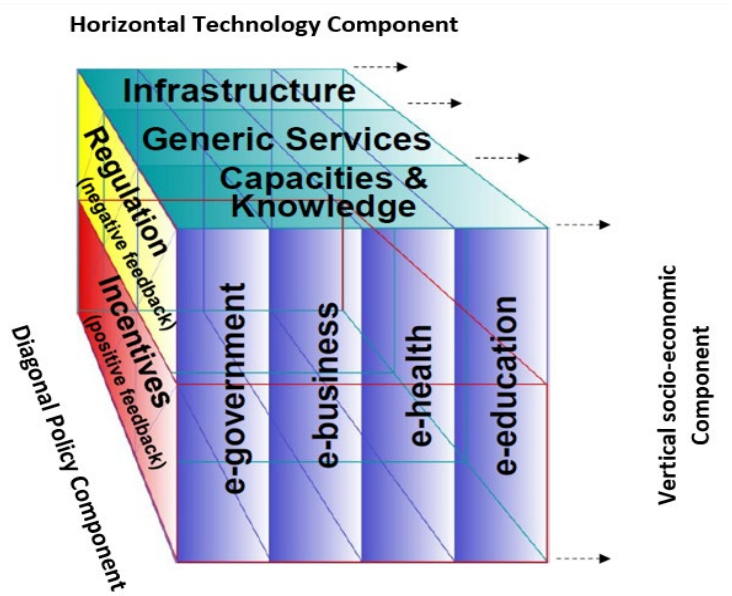


Figure 1: Modified Version of the ICT4D Cube (Hilbert, 2012)

The ICT4D-CF was developed to use as a guide for international, national and local governments, when conducting the policy process towards transitioning into an information society (Hilbert, 2012). This transition is referred to as digitalisation in this study. The three critical components of the ICT4D-CF as depicted in Figure 1 are as follows (Hilbert, 2012):

- The **Horizontal Technology Component** represents the fundamental ICT infrastructure that enables digitalisation. The ICT infrastructure includes computer software, hardware, networks and communication protocols. In computer science the ICT infrastructure is categorised in terms of the Open System Interconnection (OSI) Reference Model, which includes physical ICT and digital generic ICT services. Human skills are required to develop and implement the ICT infrastructure thus making the capacities and knowledge layer.
- The **Vertical Socio-economic Component** represents the various socioeconomic sectors to be digitalised, which are not limited e-government, e-business, e-health and e-education as shown in Figure 1.
- The **Diagonal Policy Component** represents the policy process, legislation and regulation frameworks that are supposed to promote, direct and control digitalisation.

It is the policy process (Diagonal Component in Figure 1) that plays a crucial role in digitalisation (World Economic Forum, 2020). A policy process generally involves a number of stages or cycles, from identifying policy problems and objectives, to executing, monitoring and evaluating those policy objectives (Kunyenje, 2019). The public policy process may be categorised into three broad stages of planning, execution, and evaluation (Hilbert, 2012).

The focus of this paper is on the planning stage of the governmental policy process. Government policies are crucial determinants of national socioeconomic development, hence the approach of digitalisation from a government's policy perspective. The ICT4D-CF is a useful tool in providing an overview of the critical components at play throughout the digitalisation policy process. Digitalisation is conceptualised by this paper through the ICT4D-CF in Figure 1. Ultimately, digitalisation transforms the national critical infrastructures (NCI) into NCII.

3. The National Critical Information Infrastructure

This section discusses the evolution of socioeconomic sectors from being NCI to becoming an interconnected NCII through digitalisation. An NCI transforms into an NCII once its innovations are dependent on ICT, specifically information/data and the internet. In SA, NCII is defined as:

> *"all ICT systems, data systems, databases, networks (including people, buildings, facilities and processes), that are fundamental to the effective operation of the Republic … This relates to critical services such as the economy, social services and law enforcement (inclusive of the justice system and state security)" (State Security Agency, 2015).*

From the above definition of NCII, three themes can be observed, namely: (1) the extension of ICT systems, (2) to operate national systems, (3) for socio-economic benefits. Therefore, this paper defines digitalisation as the digitisation of the National Critical Infrastructure (NCI) into an NCII resulting from the diffusion of emerging technologies across socioeconomic and political sectors. NCII describes an internet-connected strategic unit that provides essential services to the nation for the purpose of socioeconomic and political development and resilience (Runciman, 2006), through the harnessing of ICT. Twelve CIs are declared as NCII in SA and are listed, in no particular order, in Table 2.

Table 2: South African National Critical Information Infrastructure (Pillay, 2017)

Critical Infrastructures		
Manufacturing	Electricity/Energy	Police, defence and legal
Emergency services	Financial services	Telecommunications
Health services	Information Technology	Transportation and ports
Key government agencies	Liquid fuels	Water/Sanitation

On the ICT4D-CF in Figure 1, the socioeconomic sectors representing the CIs in Table 2 are prefixed with an 'e' to indicate their methods of productivity as being reliant on the internet (Hilbert, 2012). The implication of e-health or e-government, for example, is that the public may interact with computer applications through the sharing of information over the internet to access essential public services, instead of through physical interaction.

The next section presents the NCII Conceptual Framework (NCII-CF).

3.1 National Critical Information Infrastructure Conceptual Framework

Interconnectedly, Critical Infrastructures (CI) and the Critical Information Infrastructure (CII) make up the NCII. The NCII-CF is informed by a literature review and the ICT4D-CF in Figure 1. The NCII-CF, like the ICT4D-CF, consists of three critical components of technology (i.e., CII), socio-economic sectors (i.e., CI) and policy (i.e., Regulatory Framework).

CII describes an ICT infrastructure consisting of interconnected hardware, software, networking devices, equipment and services — these include computer devices, applications, operating systems, the internet and its services, and all other tangible and intangible ICT tools (Christensen et al., 2010; Hilbert, 2012; State Security Agency, 2015). These are the systems and tools that are shared within and across the various NCIs to carry out operational business activities (Runciman, 2006). As such, the emerging technologies in Table 1 are regarded as CII. Since human beings possess information in the form of intelligence and skills, they are essential in developing and operating ICT systems. As such, human beings are regarded as human capital. In the ICT4D-CF in Figure 1, human "capacities and knowledge" are similarly positioned in the horizontal technology component (Hilbert, 2012).

CIs refer to physical buildings or locations (Christensen et al., 2010) of the socioeconomic entities that are declared as NCII (see Table 2). CIs are the socioeconomic entities or organisations whose services are essential for civil development, national wealth, and national security (Sofaer et al., 2010). CIs depend on the CII for their operations, represented by the vertical socio-economic sectors component (see Figure 1).

Distinguishing between CI and CII made more sense in the previous IRs or prior to digitalisation where clear physical and digital boundaries existed. In contrast, digitalisation is essentially concerned with diminishing those boundaries, evidenced by its applications combining physical, biological and digital technologies into innovative products and services. Figure 2 presents the NCII-CF which is adapted from the ICT4D-CF in Figure 1.

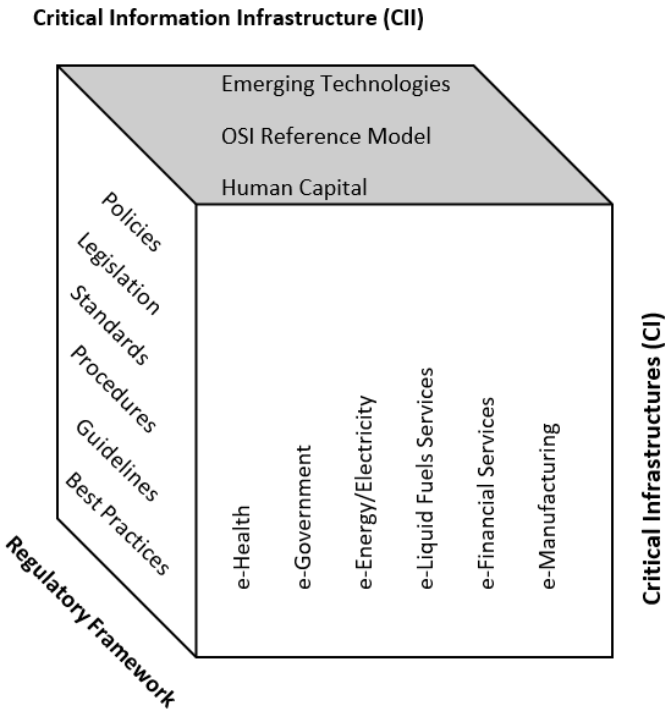


Figure 2: National Critical Information Infrastructure Conceptual Framework (adapted from Hilbert (2012))

Digitalisation implies an NCII whose production and productivity depend on emerging technologies, high interconnectivity (OSI Reference Model) and a digitally skilful workforce. Furthermore, digitalisation implies that the NCII is a borderless domain that exists and operates in cyberspace. The dependence on cyberspace of the NCII introduces significant cybersecurity concerns to governments worldwide (Kortjan, 2013; State Security Agency, 2015). This is due to the prevalence of cyberattacks on NCII (Griffiths, 2017; Pillay, 2017).

3.2 Cyberattacks on the National Critical Information Infrastructure

NCII attacks can either target CIs and thus be limited to physical damage (Christensen et al., 2010), or they can target the CII, which can result in the loss of confidentiality, integrity or availability (CIA) of the CII (State Security Agency, 2015). Digital attacks against NCII are known as cyberattacks.

According to the International Organization for Standardization and the International Electrotechnical Commission (ISO/IEC 2700:2009), an attack is "*an attempt to destroy, expose, alter, disable, steal or gain unauthorized access to or make unauthorized use of an asset*" (ISO/IEC, 2012, p. 3). If an attack and assets are in cyberspace, it is then a cyberattack. The NCII is the most targeted asset in cyberspace (Gkioulos and Chowdhury, 2021). Typical cyberattacks against NCII are cybercrime, cyberespionage, cyberterrorism, cyberwarfare or information warfare, denial of service (DoS), distributed denial of service (DDoS) and hacktivism, which are all meant to disrupt the CIA of the NCII (Christensen et al., 2010; Griffiths, 2017; Runciman, 2006; Yunos et al., 2015). The above-mentioned cyberattacks may be categorised into cyberterrorism, cyberespionage, cyberwarfare and cybercrime (Griffiths, 2017). These four categories of cyberattacks are what the Ministry of Defence Estonia (2008), and State Security Agency (2015) SA noted as new and sophisticated cyberthreats to NCII resulting from digitalisation (Ministry of Defence Estonia, 2008; State Security Agency, 2015). While cyberattacks may occur at an individual, organisational or state level, cyberattacks on the NCII are the most detrimentally impactful of all cyberattacks (Ministry of Defence Estonia, 2008).

Pillay (2017, p. 9) adds "state-to-state conflict" as a form of cyberterrorism and a cyberthreat to NCII. The assumption therefore is that cyberattacks on NCII are politically motivated and may be carried out by states against other states. This is asserted by similar studies which describe cyberterrorism as politically, socially and economically motivated cyberattacks on NCII, taking place in cyberspace and using offensive cyber tactics which are aimed at harming national government, citizens or the economy (Griffiths, 2017; Yunos et al., 2015).

Figure 3 presents a cyberterrorism conceptual framework that is meant to distinguish the six critical components that establish a cyberattack as an act of cyberterrorism (Yunos et al., 2015). A cyberterrorism conceptual

framework, such as the one in Figure 3, is significant as there are worldwide controversies and disagreements among media, social, political and academic groups concerning attributes that affirm a cyberattack as a cyberterrorism attack.

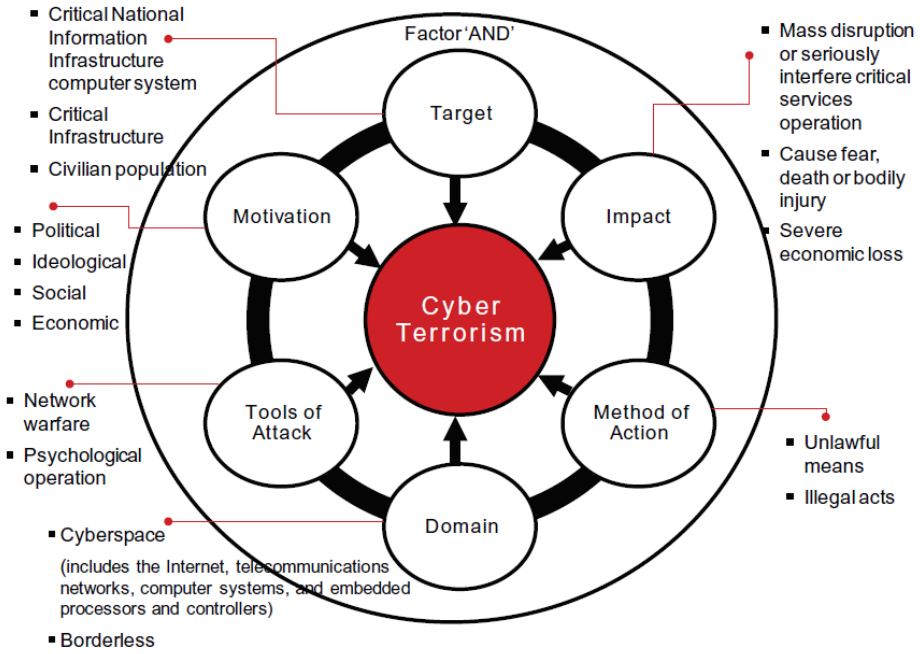


Figure 3: Cyberterrorism Conceptual Framework (Yunos et al., 2015)

Figure 3 conceptualizes six themes related to cyberterrorism concerning typical cyberattacks on NCII. The themes, presented in no particular order, are as follows (Lehto and Neittaanmäki, 2021):

- **Target (1) and Domain (2)**: Identifies components of the NCII-CF vulnerable to cyberattacks, encompassing technologies, processes, and people.
- **Method of Action (3) and Domain and Tools of Attack (4)**: Describes the ways in which the NCII may be attacked, encompassing insiders, ransomware, DDoS attacks, poor encryption, SQL injections, misconfiguration, phishing, trojans, brute force attacks, session hijacking, spoofing, cross-site scripting (XSS), etc.
- **Motivation (5)**: Explores the reasons for cyberattacks on NCII, including "Egoism, Money, Power Paralysis, and Destruction").
- **Impact (6)**: Evaluates the level of disruption that may result from such cyberattacks.

A recent study was conducted to assess the likelihood of state-to-state conflicts arising between SA and the international states it is in strategic partnerships with, which revealed SA as highly vulnerable (Van Niekerk, 2019). SA's cybersecurity approach is noted from the country's cybersecurity legislation, as being only defensive (Van Niekerk, 2019). In fact, South Africa does not have an officially documented national cybersecurity strategy (Phahlamohlaka et al., 2021). Subsequently, it has been asserted that SA has previously been a victim of cyberterrorism attributed to China and Russia. The motive of a state-to-state cyberattack is for a country to gain national power over other countries (van Vuuren et al., 2017). This paper therefore argues that the globally shared agenda of digitalisation may be detrimental to the NCII of SA given the country's vulnerable cybersecurity regulatory framework. It is granted that donors may be beneficial in assisting with the development of cybersecurity in SA (Dewa et al., 2018; G20 Osaka Leaders, 2020; Schia, 2018). However, each country will prioritise its own interests fundamentally (G20 Insights, 2017), even if it means exploiting the national cybersecurity vulnerabilities of an alliance (Griffiths, 2017; Van Niekerk, 2017).

Government has a sovereign responsibility, to ensure a cybersecure NCII as the nation transitions to digitalisation through adequate policy regulatory frameworks.

4. The National Critical Information Infrastructure Protection (Through Cybersecurity) Conceptual Framework

Digitalisation poses indispensable cybersecurity threats to the NCII of SA particularly, as it was emphasised in Section 3.2 that the country is vulnerable, has been previously targeted, and remains susceptible to politically motivated cyberattacks. National government ought to direct the implementation of relevant cybersecurity control measures through the regulatory framework (see Figure 2). Cybersecurity controls are "*means of managing [cybersecurity] risk, including policies, procedures, guidelines, practices …, which can be administrative, technical, management, or legal in nature*" (ISO/IEC, 2012, p. 4).

The South African National Standard 27032:2016 acknowledges that:

> *"… cybersecurity is one of the activities necessary for CIIP, and, at the same time, adequate protection of critical infrastructure services contributes to the basic security needs (i.e., security, reliability and availability of critical infrastructure) for achieving the goals of cybersecurity" (ISO/IEC, 2012, p. 10).*

As such, CIIP is a prioritised and indispensable responsibility of national governments across the globe. This paper suggests that national governments' cybersecurity controls are mainly strategic, such as the public policy process, national cybersecurity strategies, and international cooperation in cybersecurity. Applying the ICT4D-CF (Figure 1) and the NCII-CF (Figure 2), national governments' controls fall within the diagonal *Policy* and the *Regulatory* components, respectively. The diagonal *Policy* component generically represents the strategic controls and policy perspective of digitalisation. This paper has highlighted this component consistently as the most substantial since it directs the technology (horizontal) and socioeconomic development (vertical) components. However, it is important to acknowledge that all three components are interdependent, and none of them is exclusively effective and efficient.

Technical cybersecurity controls describe the implementation of ICT tools, systems and techniques for the purposes of maintaining the CIA of information while it is being transmitted, processed or stored in cyberspace. Applying the ICT4D-CF (Figure 1) and the NCII-CF (Figure 2), this study conceptualises technical cybersecurity controls as falling within the horizontal *Technology* and *CII* components, respectively, to secure the CIA of information assets in the NCII. Technical cybersecurity controls are referred to security domains in the ISO/IEC 27032:2012 and include information security, network security, critical information infrastructure protection (CIIP), internet security, and application security (ISO/IEC, 2012). This study focuses on CIIP which is defined as:

> *"Protecting the systems that are provided or operated by critical infrastructure providers, such as energy, telecommunication, and water departments. Critical Information Infrastructure Protection ensures that those systems and networks are protected and resilient against information security risks, network security risks, Internet security risks, as well as cybersecurity risks" (ISO/IEC, 2012, p. 11).*

The protection of NCII through cybersecurity is a national effort and combination of technical, administrative, legal, regulatory, and strategic security measures, all directed towards socioeconomic development and resilience in the advent of digitalisation. This paper proposes the NCIIP-CF in Figure 3 to conceptualise the protection of NCII through cybersecurity for national governments. The NCIIP-CF is adapted from the ICT4D-CF (Figure 1) and NCII-CF (Figure 2) and also informed by the literature that was reviewed in this study.

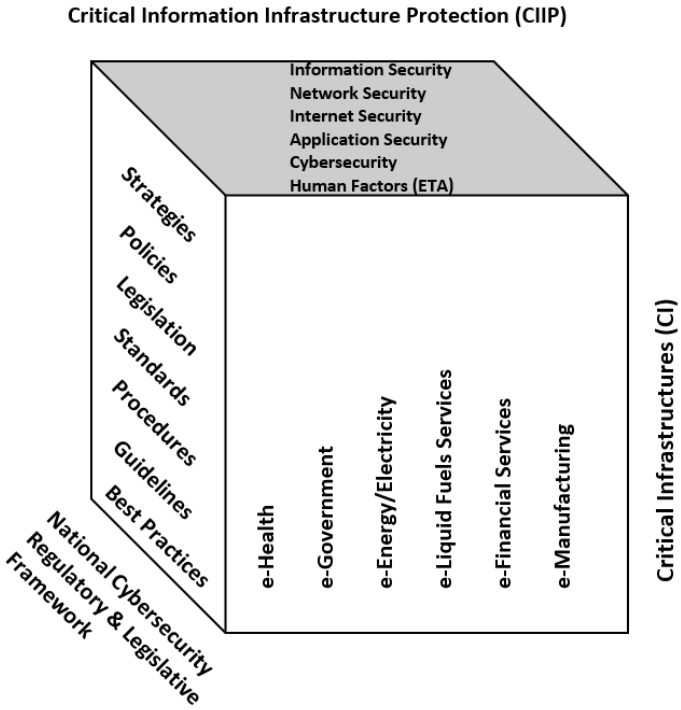


Figure 4: National Critical Information Infrastructure Protection (through Cybersecurity) Conceptual Framework (NCIIP-CF) (adapted from Hilbert (2012))

The NCIIP-CF is a logical and strategic summary and representation of digitalisation, the NCII, and national cybersecurity. It is a three-dimensional cube conceptual framework, which can be used by governments to understand the interdependent functionality of the (i.e., horizontal, vertical and diagonal) components of digitalisation, the NCII and national cybersecurity.

The (horizontal) CIIP component represents the technical cybersecurity controls. In addition to these technical cybersecurity controls are human factors which refer to the cybersecurity skills and awareness of human beings. These technical and human factors cybersecurity controls are necessary to protect each of the elements of CII.

The (vertical) CI component represents private, public and government socioeconomic sectors or organisations that will execute the national cybersecurity regulatory and legislative framework in the diagonal component of the NCIIP-CF. Generically, the vertical component represents the CIs (i.e., sectors, entities or organisations) that are subject to digitalisation.

The (diagonal) national cybersecurity regulatory and legislative framework component constitutes the strategic cybersecurity controls that national governments must lead to control digitalisation planning policy processes. This is the governments' policy perspective of digitalisation, the NCII and cybersecurity. This paper has highlighted this component consistently as the most substantial since it directs the technology (horizontal) and socioeconomic development (vertical) components. All three components of the NCIIP-CF are interdependent.

5. Conclusion

SA is a developing country whose government has aspired to socioeconomic development through the country's public policies and has acknowledged ICT to be the solution. As such, digitalisation plans are well underway in the country. This paper therefore examines digitalisation to understand the phenomenon and to determine the most effective way of implementing it, whilst minimising the cybersecurity risks that it poses.

The critical components of digitalisation include the technology *information infrastructure*, socioeconomic *critical infrastructures* and the strategic *policy* component. The locale or establishment of digitalisation is the NCII which can be categorised into three critical and interdependent components, the CII, CIs and the regulatory framework. The significance of the NCII is that it determines the well-being and survival of a nation. Therefore, disturbance to the functioning of the NCII is highly impactful and should be mitigated, primarily by national government, through national cybersecurity. The NCII of South Africa is particularly vulnerable to politically

motivated cyberattacks. Therefore, digitalisation may expose the NCII of South Africa to politically motivated state-to-state cyber-attacks that are typically committed through cyberterrorism.

As such the NCIIP-CF was proposed to conceptualise digitalisation that prioritises the protection of NCII through government's strategic national cybersecurity controls. The national cybersecurity regulatory and legislative framework component in the proposed NCIIP-CF is meant to ensure cybersecurity in cyberspace, which may result in a society and economy that is developmental, capacitated, proactive and resilient to cyberattacks.

References

Abdullah, T.A.A., Ali, W., Malebary, S., Ahmed, A.A., 2019. A Review of Cyber Security Challenges, Attacks and Solutions for Internet of Things Based Smart Home. IJCSNS International Journal of Computer Science and Network Security 19.

Bican, P.M., Brem, A., 2020. Digital Business Model, Digital Transformation, Digital Entrepreneurship: Is There A Sustainable "Digital"? Sustainability 12, 5239.

BRICS Heads of State and Government, 2018. BRICS in Africa: Collaboration for Inclusive Growth and Shared Prosperity in the 4th Industrial Revolution.

Chipidza, W., Leidner, D., 2019. A review of the ICT-enabled development literature: Towards a power parity theory of ICT4D. Journal of Strategic Information Systems.

Christensen, S., Caelli, W.J., Duncan, W.D., Georgiades, E., 2010. An achilles heel: Denial of service attacks on Australian critical information infrastructures. Information and Communications Technology Law 19, 61–85.

Cunningham, S., 2018. World economic forum and the fourth industrial revolution in South Africa. Trade & Industrial Policy Strategies 30.

Department of Science and Technology, 2019. White paper on science, technology and innovation. Department of Science and Technology 1–70.

Dewa, M.T., Adams, D.Q., Tendayi, T.G., Nyanga, L., Gxamza, M., Ganduri, L., 2018. Industry 4.0: A myth or a reality in South Africa? Proceedings of the 29th South African Institute of Industrial Engineers Conference 575–590.

G20 Insights, 2017. G20 and Africa - Ready for a Steady Partnership ? [WWW Document]. URL www.G20-insights.org (accessed 5.21.20).

G20 Osaka Leaders, 2020. G20 Osaka Leaders' Declaration 1–13.

Gkioulos, V., Chowdhury, N., 2021. Cyber security training for critical infrastructure protection: A literature review. Comput Sci Rev.

Griffiths, J.L., 2017. Cyber security as an emerging challenge to South Africa National Security 88.

Hilbert, M., 2012. Toward a Conceptual Framework for ICT for Development: Lessons Learned from the Cube Framework Used in Latin America (English). Information Technologies & International Development 8, 243–259.

ISO/IEC, 2012. ISO 27032 Information technology - Security techniques - Guidelines for cybersecurity. International Organization for Standardization 50.

Kortjan, N., 2013. A Cyber Security Awareness and Education Framework for South Africa 206.

Kunyenje, G., 2019. Influence of External Actors on National Information and Communications Technology Policy Formulation in Developing Countries: Case of Malawi. University of Cape Town.

Laufs, J., Borrion, H., Bradford, B., 2020. Security and the smart city: A systematic review. Sustain Cities Soc 55.

Lehto, M., Neittaanmäki, P., 2021. Critical Infrastructure Protection Cyber Security. In: Lehto, M., Neittaanmaki, P. (Eds.), Computational Methods in Applied Sciences. Springer Nature Switzerland AG, Cham, pp. 3–36.

Ministry of Defence Estonia, 2008. Cyber Security Strategy Ministry of Defence.

Mitchell, C.J., 2020. The impact of quantum computing on real-world security: A 5G case study. Comput Secur 93, 1–11.

Ordóñez, A., 2015. Impact of Information Society Research in the Global South, A New Set of Questions: ICT4D Research and Policy. Springer Singapore, Singapore.

Perez, C., 1985. Microelectronics, long waves and world structural change: New perspectives for developing countries. World Dev 13, 441–463.

Phahlamohlaka, J., Theron, J., Aschmann, M.J., 2021. National Cybersecurity Implementation in South Africa : The Conundrum Question.

Pillay, K., 2017. National Critical Information Infrastructure Protection Centre. NCIIPC Newsletter 1, 40.

Pillay, K., Ori, A., Merkofer, P., 2016. Industry 4.0: Is Africa ready for digital transformation?, Delloite South Africa.

Runciman, B., 2006. who protects the information our critical infrastructure relies on? [WWW Document]. URL http://0-search.ebscohost.com.wam.seals.ac.za/login.aspx?direct=true&db=bsu&AN=24607744&site=ehost-live&scope=site

Schia, N.N., 2018. The cyber frontier and digital pitfalls in the Global South. Third World Q 39, 821–837.

Schumpeter, J.A., 1939. Business cycles: A theoretical, historical and statistical analysis of the capitalist process, Abridged. ed, McGraw-Hill Book Company. McGraw-Hill Book Company, New York Toronto London.

Schwab, K., 2016. The Fourth Industrial Revolution. World Economic Forum 91–93 route de la Capite CH-1223 Cologny/Geneva Switzerland, Geneva, Switzerland.

Schwab, K., 2020. Now is the time for a "great reset" of capitalism | World Economic Forum. World Economic Forum.

Sofaer, A., Clark, D., Diffie, W., 2010. Cyber security and international agreements. In: Proceedings of a Workshop on Deterring Cyberattacks. pp. 1–82.

State Security Agency, 2015. National Cybersecurity Policy Framework For South Africa 66–95.

Sutherland, E., 2017. Governance of Cybersecurity – The Case of South Africa. The African Journal of Information and Communication 83–112.

Thun, S., Kamsvåg, P.F., Kløve, B., Seim, E.A., Torvatn, H.Y., 2019. Industry 4.0: Whose revolution? The digitalization of manufacturing work processes. Nordic Journal of Working Life Studies 9, 39–57.

Van Niekerk, B., 2017. An Analysis of Cyber-Incidents in South Africa. The African Journal of Information and Communication 113–132.

Van Niekerk, B., 2019. The cyber security dilemma: A South African perspective. 14th International Conference on Cyber Warfare and Security, ICCWS 2019 441–447.

van Vuuren, J., Leenen, L., Plint, G., Zaaiman, J., Phahlamohlaka, J., 2017. Formulating the Building Blocks for National Cyberpower. International Journal of Cyber Warfare and Terrorism 7, 16–28.

Victoria, A., 2019. Cybersecurity on 5G Technology.

Wang, A.X., Hou, A., Warren, B., Popova, I., Shelepov, A., Sakharov, A., Ignatov, A., 2019. 2018 BRICS Johannesburg Summit Interim Compliance Report Prepared by 1–101.

World Economic Forum, 2020. Energy Policy Lighthouses: The Little Green Book.

Yunos, Z., Ahmad, R., Mohd Sabri, N.A., 2015. A Qualitative Analysis for Evaluating a Cyber Terrorism Framework in Malaysia. Information Security Journal 24, 15–23

Intelligence Agencies' Move to the Cloud: Challenges and Opportunities

Karin Säberg[1] and Gazmend Huskaj[1,2]
[1]Department of Computer and Systems Sciences, Kista, Sweden
[2]Geneva Centre for Security Policy, Geneva, Switzerland

karin.saberg@gmail.com
g.huskaj@gcsp.ch

Abstract: The purpose of this research is to discover more about the challenges and opportunities faced by intelligence agencies wishing to move their data to the cloud. Intelligence agencies collect and process enormous amounts of data and information and need the tools to do so. Two intelligence communities have moved to the cloud to face these issues but there is little scientific knowledge about moving an intelligence agency's data to the cloud. No research on the topic could be found and this study aims to fill part of that gap by using a case study research strategy and interviews with experts in the field. A literature review was completed to understand previously identified challenges when adopting cloud and was used to create two sets of interview questions. Five interviews were conducted, and a thematic analysis done resulting in fourteen themes. The themes revealed that there are many challenges with laws and regulations being the biggest one, while the opportunities brought by a cloud solution are the processing and analysis of data, and information sharing.

Keywords: Intelligence, Intelligence agency, Cloud, Cloud computing, Information security

1. Introduction

Intelligence agencies use multiple intelligence-collection methods: Human Intelligence, Signals Intelligence, Imagery Intelligence, Measurement and Signatures Intelligence, and Open-Source Intelligence (OSINT) (FBI, n.d.; MI5, n.d.; SÄPO, 2022). Of these, open-source intelligence is of growing interest, especially in today's highly digitalised world. The term open-source intelligence is relatively recent, having been coined in the late 80s by the US military (Schaurer & Störger, 2013). It concerns collecting data from "publicly available information, as well as other unclassified information that has limited public distribution or access" (NATO, 2013). SÄPO (2022) exemplifies OSINT using the internet, newspapers, radio and tv.

Intelligence agencies need to discover, access, and share critical all-source, multi-fused intelligence in a timely, accurate and actionable manner (Palfy, 2015; Schaurer & Störger, 2013). To accomplish their goals, intelligence agencies face two challenges: the "handling" (discovery, access, analysis, sharing) of the data and related security. To tackle these issues, the Central Intelligence Agency (CIA) employed cloud solutions for all 17 United States intelligence agencies (Konkel, 2014).

The deal was for Amazon Web Services (AWS) to provide this solution for ten years. The CIA reached a new agreement with AWS, Microsoft, Google, Oracle, and IBM (Konkel, 2020). Little is known about the specifics and requirements that supported the move to the cloud, both in 2014 and 2020, other than a statement from the associate deputy director at CIA's Digital Innovation Directorate, Sean Roche:

> *"[…] the cloud on its weakest day is more secure than a client service solution", and government organisations are desperately trying to upgrade legacy information technology (IT) systems (Corrigan, 2018).*

The scientific literature reveals scarce research on the topic, covering the US and UK's intelligence communities (Konkel, 2014; Konkel, 2020; Warrell & Fildes, 2021). None provide in-depth information about an intelligence agency's decision processes, challenges, (security) requirements and practical solutions. The most extensive study on the subject is a paper by Horlings (2022) discussing using computer science tools as part of the intelligence process. However, even she does not mention cloud solutions despite the verbatim mention of "scalable architecture for efficient storage, manipulation, and analysis" (Horlings, 2022, p. 13), a description that corresponds to cloud computing.

As such, there are issues with how to approach information security and the potential resistance to moving to the cloud. Facing these issues and understanding what intelligence agencies need to consider adopting cloud solutions requires additional research.

Specifically, this study considers the following research questions:

- What challenges do intelligence agencies face when adopting cloud solutions?

- What opportunities does a cloud solution bring to an intelligence agency?

This study aims to discover intelligence agencies' challenges in moving to the cloud. However, this research will also explore opportunities in using the cloud, which would also likely bring opportunities, such as tools and solutions benefitting their work.

2. Methods and Materials

Case studies aim to understand social phenomena where it is of interest to study a specific thing in-depth by using different data and research methods (Denscombe, 2014; Yin, 2018). This study is both exploratory and explanatory, and case studies lend themselves well for both types (Yin, 2018). This study's aims and research question is phrased using what but truly explore the reasoning and whys behind intelligence agencies moving to the cloud. Case studies work well to explain and understand how different facets affect the case (Denscombe, 2014), and to explore current events (Yin, 2018), in this case the lack of research concerning the intelligence community and the cloud.

Using a case study research strategy with interviews as data collection method five experts in the field of intelligence agencies and the cloud were interviewed. Runeson and Höst (2009) proposes a five-step process when conducting a case study which was followed. The steps are as follows:

1. Case study design – plan the study and research.
2. Preparation for data collection – literature review leading to conceptualising and interviews based on these. Reviewed Denscombe (2014) and Yin (2018) for case study design.
3. Collecting evidence – semi-structed interviews circa 90 minutes long with experts in the field.
4. Analysis of collected data – thematic analysis according to Braun and Clarke (2022) and Denscombe (2014).
5. Reporting – writing the paper.

2.1 Data Collection

Purposive sampling and semi-structured interviews were used to collect data from experts in the field chosen based on their knowledge and experience in the intelligence community and their availability to participate. The experts have at least 10 years of experience, and varying degrees of experience with cloud solutions, either in research for intelligence purposes, procuring, or developing cloud solutions. Four of the interviews were performed in person and one using Zoom, and two of the interviews were recorded with consent. The interview questions cover two main themes: intelligence and technological aspects of moving an intelligence agency's data to the cloud. The two sets of interview questions were based on the concepts found in table 1 and 2, with some overlap between the two.

Experts 1, 2 and 5 were asked the intelligence questions, and experts 3 and 4 the technical questions. As the literature review revealed that technical challenges when moving an organisation's data and processes to the cloud are largely the same or similar independent on the user and purpose of the cloud, while challenges faced by intelligence agencies are nearly completely unaddressed. Thus, the need to understand the intelligence community's view on challenges on cloud solution adoption is in more dire need than addressing an already well-researched area. As such one more interview was performed to explore this specific aspect.

2.2 Data Analysis

A qualitative thematic data analysis was performed on the data, following a five-step process proposed by Denscombe (2014, pp. 247-248), and using Braun and Clarke (2022) for a deeper understanding.

The five-step process by Denscombe (2014) is described in Table 3.

Table 1: Overview of Data Analysis Process and its Application

Step	Step Five-stage data analysis	Application in this study
1	Data preparation – Transcribe the text, cataloguing or loading to software,	Notes from interviews are transcribed
2	Initial exploration of the data – Look for themes, add notes and write memos,	1st (initial) iteration
3	Analysis of the data – Code the data, group the data into categories, compare categories, further group them into concepts,	Two iterations planned
4	Presentation and display of the data – Written interpretation of the findings, illustrate	Present final iteration

Step	Step Five-stage data analysis	Application in this study
	with quotes, use visual means of presentation,	themes
5	Validation of the data – Data and method triangulation, alternative explanations comparison.	Compare to previous research etc.

Using Taguette, an open-source tagging tool, to complete the first iteration of coding, 151 codes were identified. The codes, with the relevant text, was printed and sorted into 24 categories, during this process one code was discarded as irrelevant. The 24 categories and 150 codes were next entered into FreeMind, a mind mapping tool, and sorted into themes. A total of 14 main themes were identified during this process.

3. Results

The results of the thematic analysis, depicted in Figure 1, show 14 main themes which will be explained here. While some are in line with previous research, others are not. The themes identified, in no specified order, are as follows:

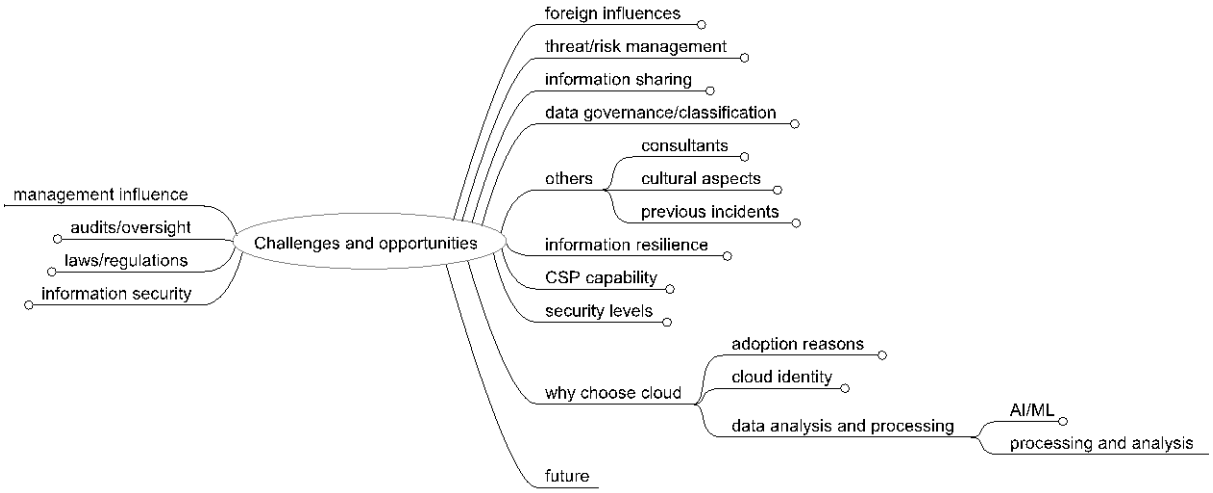


Figure 1: Thematic Analysis Results, with some sub-themes. Note: CSP = Cloud service provider, *Security levels* is renamed to *security clearance level* in the text

3.1 Management Influence

This theme explores management's role and influence on cloud adoption and usage. It covers Service Level Agreements (SLAs) procurement, requirements, budgeting, and resource allocation. The experts we interviewed were divided on SLAs. While one expert believed in existing expertise, two others noted for improved "contractual capacity." These experts felt that those managing SLAs and contracts often lack the legal or technical competence.

Management should consider internal and external requirements and understand how to measure the usage requirements for cloud solutions. Initial resource allocation and budget for cloud transition was a highlighted concern.

A significant issue is the management's lack of knowledge about cloud solutions, their internal workings, and what using one means for an intelligence agency. Management must close their knowledge gap before real cloud solution adoption can start. This is currently a hurdle, and the experts noted that change comes from the top.

3.2 Audits/Oversight

The necessity for audits remains unchanged from existing solutions. However, a cloud solution could likely make oversight easier. Because both audits and oversight trace actions in the system, it also relates to the contractual and SLA-aspects discussed under the management influence theme. SLAs and contracts should support audits and logging tools to a sufficient degree.

3.3 Laws/Regulations

Laws and regulations often lag behind the technological and digital evolution, and both national laws and regulations, as well as EU law like GDPR must be adhered to. The experts noted that Swedish government agencies are often very autonomous and may have different regulations regarding cloud usage, further complicating the issue. Also, regulations and laws regarding secret and top secret classified information are unclear, and does not support cloud usage at all, meaning that the full potential of a cloud cannot be utilised if not all necessary information can be put in a cloud solution. According to the experts, even though changes to legislation is being made, it is slow and is not keeping up with the needs.

3.4 Information Security

This theme covers many aspects of information security, like best practices, processes, information security principles, and risks with new technologies. The experts agree following best practice recommendations and staying up to date with current technologies. Encryption was mentioned as a possible solution to manage information security risks, which also requires relevant processes to manage encryption.

Information security principles include zero trust environments, two-men rules, and least-privilege access. This were reoccurring practices mentioned by the experts as ways to ensure security.

Risks include new technologies and methodologies: they should not be used until tried and tested. Additional risks include the cloud-specific data segregation issue. Although a benefit of the cloud is the use of shared resources, this is not always good for an intelligence agency: it needs to be strictly managed through data segregation to ensure the security and secrecy of information. This also limits cloud usage possibilities.

3.5 Foreign Influences

Concerns were raised adopting and using cloud solutions: foreign actors can leverage these solutions to increase their influence. Mostly US government and US companies were in focus that the US government can demand access to US company data, and by extension, their customers. One expert expressed a concern of government's right to access data versus their actual ability to do so.

There are further issues with more entities adoption cloud solutions putting pressure on those who have not to be able to keep up in the international intelligence community. Four of the experts also expressed concerns about being denied access to data if a foreign cloud service provider were to be used. However, the main concern in this theme is the legal issues present in using a US cloud service provider and the security issues this brings.

3.6 Threat/Risk Management

Intelligence agencies need to take risks and balance threats and opportunities, a reoccurring topic. The experts argued that it's impossible to have a zero-risk approach using a cloud solution and vulnerabilities do exist. Therefore, to utilise these Internet-based tools, threats and opportunities must be well-balanced.

3.7 Information Sharing

Information sharing within a country's intelligence community, and between different countries' intelligence communities, is of importance. The experts indicated that information sharing would be easier using cloud solutions and may likely become a necessity in the future. Furthermore, intelligence reports are being shared today, but in the future, by using the cloud, data rather than reports may be shared. Providing access to data requires appropriate security controls to ensure that the right people have access to the right data at the right time. Cloud solutions can provide the tools to control information sharing and who has access to what, and reduce the risk of missing critical points and data spreading.

3.8 Data Governance/Classification

Governance, and a unified and same classification system is needed if multiple agencies were to use the same cloud solution. The interviews revealed that intelligence agencies and communities use different security and data classification schemes: different classifications may be of the same kind, or the same data may be classified differently. Not only that, but the same information may have different data security levels. This poses challenges to information sharing as well as using the same cloud.

3.9 Information Resilience

Information resilience was noted by one of the experts as an important factor for intelligence agency's use of the cloud, especially the military. Information resilience is about having the capability to continue working during crisis and war. The cloud can be an important tool for this. Unbound by physical laces that can be destroyed, military capability is ensured no matter where people are and their situation. This also includes sovereignty, a nation's ability to control its digital tools and data (Couture & Toupin, 2019; Floridi, 2020).

3.10 Cloud Service Provider Capability

Cloud service provider capability is important. In particular, the ability to deliver on data and information security, and redundancy. A lot of resources, especially staff, are needed to manage self-hosting solutions. The experts noted that if the US, with far more resources than most governments, opted for a cloud solution rather than self-hosting, then this indicates that the needs and requirements are far too cumbersome for smaller countries. Here, cloud service providers have resources to handle any potential issues. However, cloud service providers have different services and locking oneself to cloud-specific solutions can turn into issues further down the road, which is why the capabilities and services of cloud providers should be considered carefully.

3.11 Security Clearance Levels

Organisations handling data, information and intelligence related to national security value security clearance levels. Therefore, when choosing and planning for a cloud solution, those responsible should implement, plan, and consider how to manage security clearance levels for data and information. This becomes even more important if multiple intelligence agencies and government bodies share intelligence. The aim is to classify data and information correctly and to address some of the personnel security clearance. However, security clearance is (presumably) already well practiced and is not an issue specific to cloud-solutions.

3.12 Why Choose the Cloud?

Here, the reasons, benefits and challenges are presented. The findings show that intelligence agencies currently use old and slow tools. The reasons to adopt a cloud solution include, automation, speeding up work by utilising more efficient tools, sharing systems, resources, scalability, data format, different deployment models (community, hybrid, private, and public clouds), and similar benefits. Choosing a cloud solution provides important technical advantages, the most important being data processing and analysis. The experts noted that information overload has always been an issue for intelligence agency. Utilising a cloud to automate and speed up data processing and analysis can mitigate information overload.

The experts raised their concerns on using machine learning (ML) and artificial intelligence (AI) to make decisions without proper human oversight. Although ML and AI speed up data processing and analysis, human oversight and checks in place are needed to reduce these concerns.

3.13 Future

This theme is unique compared to the rest. While the other themes touch upon more specific challenges and opportunities, the findings suggest that the cloud is the future for Intelligence communities, partly because of the themes already covered above, but more so the opportunities the cloud brings organisations, especially one that handles massive amounts of data like an intelligence agency. The need to process and analyse vast amounts of data requires the appropriate tools, and cloud computing is such as tool. And this is nothing new.

One expert noted that Google, for example, is already using the cloud for intelligence purposes. They process and analyse their data through cloud computing to predict our needs to direct advertisements. What is this, if not intelligence? And it is correct. Google uses the cloud to produce intelligence about their customers to direct advertisements to them: the cloud is part of an intelligence process. Therefore, this is already happening every minute and day and is a possibility with the cloud, an argument to consider well before dismissing it.

3.14 Others

This theme combines three topics: consultants, cultural aspects, and previous incidents. The experts mentioned consultants as sources of risk and opportunities. Third-party and independent actors can bring valuable services such as audits. Cultural aspects in an organisation were possible challenges towards cloud adoption, but some groups support the cloud and would like to see it used.

A potential challenge to cloud adoption was previous incidents; there have been some cases where Swedish government agencies have attempted to adopt cloud solutions which have failed. Attempts that have failed due to mismanagement and a "you get what you pay for" attitude can work against future cloud adoption attempts and proposals but can, and should, be learnt from to understand what went wrong.

4. Discussion

The findings of this study show that the challenges faced by intelligence agencies moving to the cloud are many, one of the biggest issues is information security and cloud service provider capability. The demanding requirements on information security to ensure that the data in the cloud is only accessible by those authorised means that a CSP need to prove their capability to provide these services. The fact that many, if not most, of the big CSPs are US based and subject to US legalisation was also one major issue brought up. While encryption was proposed as one solution to mitigate this it also requires trust that the CSP delivers the encryption and services as promised, if this cannot be guaranteed it is unlikely that an intelligence agency would even consider choosing a US-based CSP. While previous research (Alouffi et al., 2021; Balani & Varol, 2020; Bhajantri & Mujawar, 2019; Butt et al., 2022; Wulf et al., 2019) touched on the data security of information security it did so in more general terms and discussed data security broadly rather than its components which, considering the selection of previous research for this study is logical, as there are plenty of studies that discuss specific data security controls. CSP's capabilities such as the lack of interoperability and portability was found in previous research (Alouffi et al., 2021; El-Gazzar, 2014; Jones, 2015; Verma & Adhikari, 2020; Zwattendorfer et al., 2013), while the other challenges found by this study such as CSPs' ability to provide information security solutions are new.

The issues brought by US legalisation are tied to the foreign influences theme where there is both the challenge of foreign legalisation, but also how their use of the cloud currently means they are getting ahead in the intelligence game. The previous research regarding foreign influence does not mention the same concerns here but rather that they are already using cloud solutions and can thus reap the benefits of it already. To be able to keep up the adoption of cloud solutions is likely to be needed, and this is expressed multiple times by the experts interviewed.

While audits were noted as a possible challenge in previous research (El-Gazzar, 2014; Tweneboah-Koduah et al., 2014; Wulf et al., 2019; Zwattendorfer et al., 2013) this was not the case here, rather the possibility for improved oversight and log management (part of audits and tracking actions) was the main point.

One finding of interest is the management influence theme. Management is, in previous research (El-Gazzar et al., 2016; Horlings, 2022; Jones, 2015; Liang et al., 2017; Oliveira et al., 2014; Porrawatpreyakorn et al., 2019), commonly brought up as a supporting and driving factor in adopting new technology and solutions and it is usually the education and expertise of the employees that can be noted as a future and further challenge. This study found that while some employees may lack knowledge and expertise in using the cloud it is the management's lack of knowledge that is the big hinder. They are not aware of its many uses, abilities, and risks, and are the ones who need to learn more.

Employees are, in private companies, more likely than not to use cloud solutions (Eurostat, 2021; Galov, 2023; Lukehart, 2022) and thus more likely to possess knowledge about its use. Management, especially in intelligence communities, appears to have been using old and outdated technology for a long time and thus lack the necessary experience with cloud solutions. The findings of this study are that management is more likely to need to learn about the benefits, uses, and nuances of cloud computing rather than employees. This is not to say they completely lack this knowledge, they are learning and slowly understanding the need for the cloud.

Another challenge noted, and supported by previous research, is the laws/regulation theme. Previous research (El-Gazzar, 2014; Horlings, 2022) noted how laws and regulations have not kept up with the technological advances and need to be updated and revised to better reflect the current digital environment. The findings in this study support this, and that intelligence agencies often face even more complex issues as, in the case of Sweden, the legalisation does not support moving top secret and secret classified information and data to the cloud, and in some cases outside the borders at all. Laws and regulations need to be updated and allow for moving highly classified information outside the borders for clouds to become a viable option. If the limit of keeping the information within the borders stays it brings up the issue of information resilience, wherein the continued operability and capability would be hindered if the cloud premises were to be placed within a country. By limiting where the cloud is physically placed one of the major benefits of using a cloud solution is

negated, the potential to have the data stored far away protects it in the case of local crisis, both natural and manmade.

While many of the themes found presented challenges, they are of the type to be considered when choosing what and how a cloud should be implemented, such as planning for data governance/classification, information sharing, security clearance levels, and threat/risk management. These are considerations to mind for any new system to be implemented, especially for organisations with high-security demands. While some of these themes are supported by previous research threat/risk management was not one of these. However, it should be part of any security process to consider the threats and risks when adopting new solutions and such research can be referenced if needed.

As Lahneman (2010) noted information sharing will be required and the results support this, Horlings' (2022) findings regarding the difficulties in information sharing can, by using a cloud solution, be mitigated. Palfy (2015) and Horlings (2022) both noted the need for data governance which the findings of this study support. No previous research regarding security clearance levels was found as part of the literature review and is a new finding.

The others theme presents some considerations to be made regarding previous incidents and how to handle cultural aspects which also should be noted. The cultural aspects of this theme is supported by earlier findings (Horlings, 2022; Koç et al., 2022; Liang et al., 2017), while the previous incidents are not and should be considered new knowledge to the field.

The themes that can be classified as opportunistic in nature, to answer the question What opportunities does a cloud solution bring to an intelligence agency? are: why choose cloud and future. However, the themes information sharing, information resilience, and audits/oversight also include many opportunities.

Why choose cloud and future both note reasons as to why one should use the cloud, especially the analytical and processing benefits. The opportunities that can be found in the other three themes are perhaps less obvious but nonetheless as important. The potential to share information easily and control who has access to what while ensuring secure sharing cannot be dismissed, just as the ability for improved oversight cannot be overlooked. Improved oversight of data and information access was noted by the experts to be one big potential change and benefit to using a cloud solution. This would allow for easier tracking of who has accessed and viewed what meaning that information security can be improved.

The information resilience theme is one of the most opportunistic in nature, it would allow for continued operationality and capability in case of crisis. If the data and systems were cloud-based, then personnel would be able to keep working no matter what happens to their physical workplace. Cloud solutions for military use would likely benefit the most of this, even if the country of operations were to be attacked, they could continue their intended operations. During the interviews the war in Ukraine was brought up as an example of how the cloud can support and benefit a country at war, but the details are outside the scope of this study. Furthermore, information resilience was not one of the concepts found during the previous research review and thus provided some interesting insight into intelligence-specific challenges not found in other organisations.

5. Conclusions

This study aimed to answer the following research question: What challenges do intelligence agencies face when adopting cloud solutions? And the following sub-question: What opportunities does a cloud solution bring to an intelligence agency?

The main finding of this study is that laws and regulations are the biggest challenge for intelligence agencies moving to the cloud. This challenge is beyond the direct control of the intelligence community, as it depends on the legal frameworks of different countries and jurisdictions. Local laws may restrict the use of cloud services, while foreign laws, especially those of the US, may pose a threat to the security and privacy of the data stored in the cloud.

The main benefit of a cloud solution for intelligence agencies is the enhanced analytical and processing capabilities, as well as the potential for information sharing and collaboration. These advantages may outweigh the risks of cloud adoption, if the risks are properly managed and mitigated. However, this requires a high level of technical expertise and organisational readiness from the intelligence agencies and their partners.

This study contributes to the literature on intelligence agencies that wish to move to the cloud by exploring the specific challenges and opportunities faced by intelligence agencies in this domain. It also highlights the gap between the technological and legal requirements of cloud adoption, which is often caused by the lack of appropriate technical knowledge among the lawmakers and regulators.

A possible direction for future research is to investigate how this gap can be bridged, and how the legal frameworks can be updated and harmonised to facilitate the use of cloud services by intelligence agencies, while ensuring the protection of their data and interests.

References

Alouffi, B., Hasnain, M., Alharbi, A., Alosaimi, W., Alyami, H., & Ayaz, M. (2021). A systematic literature review on cloud computing security: Threats and mitigation strategies. IEEE Access, 9, 57792–57807. https://doi.org/10.1109/ACCESS.2021.3073203.

Balani, Z., & Varol, H. (2020). Cloud computing security challenges and threats. In 2020 8th International Symposium on Digital Forensics and Security (ISDFS) (pp. 1–4). https://doi.org/10.1109/ISDFS49300.2020.9116266.

Bhajantri, L. B., & Mujawar, T. (2019). A survey of cloud computing security challenges, issues and their countermeasures. In Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019) (pp. 376–380). https://doi.org/10.1109/I-SMAC47947.2019.9032545.

Braun, V., & Clarke, V. (2022). Thematic analysis. SAGE.

Butt, U. A., Amin, R., Mehmood, M., Aldabbas, H., Alharbi, M. T., & Albaqami, N. (2022). Cloud security threats and solutions: A survey. In Wireless Personal Communications. Springer. https://doi.org/10.1007/s11277-022-09960-z.

Corrigan, J. (2018). CIA official: Cloud is more secure than old tech, less "soul-crushing." Nextgov. https://www.nextgov.com/it-modernization/2018/06/cia-official-cloud-more-secure-old-tech-less-soul-crushing/149211/.

Couture, S., & Toupin, S. (2019). What does the notion of "sovereignty" mean when referring to the digital? New Media and Society, 21(10), 2305–2322. https://doi.org/10.1177/1461444819865984.

Denscombe, M. (2014). The good research guide: For small-scale social research projects (5th ed.). Open University Press.

El-Gazzar, R. (2014). An overview of cloud computing adoption challenges in the Norwegian context. In Proceedings - 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing, UCC 2014 (pp. 412–418). https://doi.org/10.1109/UCC.2014.52.

El-Gazzar, R., Hustad, E., & Olsen, D. H. (2016). Understanding cloud computing adoption issues: A Delphi study approach. Journal of Systems and Software, 118, 64–84. https://doi.org/10.1016/J.JSS.2016.04.061.

Eurostat. (2021). Cloud computing - statistics on the use by enterprises - Statistics Explained. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Cloud_computing_-_statistics_on_the_use_by_enterprises.

FBI. (n.d.). Types of intelligence collection - Intelligence studies. LibGuides at Naval War College. Retrieved January 30, 2023, from https://usnwc.libguides.com/c.php?g=494120&p=3381426.

Floridi, L. (2020). The fight for digital sovereignty: What it is, and why it matters, especially for the EU. Philosophy and Technology, 33(3), 369–378. https://doi.org/10.1007/s13347-020-00423-6.

Galov, N. (2023). Cloud adoption statistics for 2023. Web Tribunal. https://webtribunal.net/blog/cloud-adoption-statistics/.

Horlings, T. (2022). Dealing with data: Coming to grips with the Information Age in Intelligence Studies journals. Intelligence and National Security. https://doi.org/10.1080/02684527.2022.2104932.

Jones, S. (2015). Cloud computing procurement and implementation: Lessons learnt from a United Kingdom case study. International Journal of Information Management, 35(6), 712–716. https://doi.org/10.1016/j.ijinfomgt.2015.07.007.

Koç, B., Şener, U., & Eren, E. P. (2022). Determinative factors of cloud computing adoption in government organizations. In 2022 3rd International Informatics and Software Engineering Conference (IISEC). https://doi.org/10.1109/IISEC56263.2022.9998286.

Konkel, F. (2014). The details about the CIA's deal with Amazon. The Atlantic. https://www.theatlantic.com/technology/archive/2014/07/the-details-about-the-cias-deal-with-amazon/374632/.

Konkel, F. (2020). CIA awards secret multibillion-dollar cloud contract. Nextgov. https://www.nextgov.com/it-modernization/2020/11/exclusive-cia-awards-secret-multibillion-dollar-cloud-contract/170227/.

Lahneman, W. J. (2010). The need for a new intelligence paradigm. International Journal of Intelligence and CounterIntelligence, 23(2), 201–225. https://doi.org/10.1080/08850600903565589.

Liang, Y., Qi, G., Wei, K., & Chen, J. (2017). Exploring the determinant and influence mechanism of e-Government cloud adoption in government agencies in China. Government Information Quarterly, 34(3), 481–495. https://doi.org/10.1016/j.giq.2017.06.002.

Lukehart, M. (2022). Cloud computing statistics. Parachute. https://parachute.cloud/cloud-computing-statistics/.

MI5. (n.d.). Gathering intelligence. MI5 - The Security Service. Retrieved January 30, 2023, from https://www.mi5.gov.uk/gathering-intelligence.

NATO. (2013). NATOTerm. https://nso.nato.int/natoterm/content/nato/pages/home.html?lg=en.

Oliveira, T., Thomas, M., & Espadanal, M. (2014). Assessing the determinants of cloud computing adoption: An analysis of the manufacturing and services sectors. Information and Management, 51(5), 497–510. https://doi.org/10.1016/j.im.2014.03.006.

Palfy, A. (2015). Bridging the gap between collection and analysis: Intelligence information processing and data governance. International Journal of Intelligence and CounterIntelligence, 28(2), 365–376. https://doi.org/10.1080/08850607.2015.992761.

Porrawatpreyakorn, N., Tangprasert, S., Nuchitprasitchai, S., Chaipunyathat, A., & Viriyapant, K. (2019). Understanding key enablers of cloud computing adoption and acceptance over time. In 2019 Research, Invention, and Innovation Congress (RI2C).

Runeson, P., & Höst, M. (2009). Guidelines for conducting and reporting case study research in software engineering. Empirical Software Engineering, 14(2), 131–164. https://doi.org/10.1007/s10664-008-9102-8.

SÄPO. (2022). Underrättelsearbete. Säkerhetspolisen. https://sakerhetspolisen.se/verksamheten/underrattelsearbete.html.

Schaurer, F., & Störger, J. (2013). The evolution of open source intelligence (OSINT). The Intelligencer Journal of U.S. Intelligence Studies, 19(3), 53–56. Retrieved from https://www.afio.com/publications/Schauer_Storger_Evo_of_OSINT_WINTERSPRING2013.pdf.

Tweneboah-Koduah, S., Endicott-Popovsky, B., & Tsetse, A. (2014). Barriers to government cloud adoption. International Journal of Managing Information Technology, 6(3), 1–16. https://doi.org/10.5121/ijmit.2014.6301.

Verma, G., & Adhikari, S. (2020). Cloud computing security issues: A stakeholder's perspective. SN Computer Science, 1(6). https://doi.org/10.1007/s42979-020-00353-2.

Warrell, H., & Fildes, N. (2021, October 26). Amazon strikes deal with UK spy agencies to host top secret data. Financial Times. https://www.proquest.com/docview/2601746909/citation/EDD5BD10ABC2411FPQ/4?accountid=38978.

Wulf, F., Strahringer, S., & Westner, M. (2019). Information security risks, benefits, and mitigation measures in cloud sourcing. In Proceedings - 21st IEEE Conference on Business Informatics, CBI 2019 (Vol. 1, pp. 258–267). https://doi.org/10.1109/CBI.2019.00036.

Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE.

Zwattendorfer, B., Stranacher, K., Tauber, A., & Reichstädter, P. (2013). Cloud computing in E-government across Europe: a comparison. In Technology-Enabled Innovation for Democracy, Government and Governance: Second Joint International Conference on Electronic Government and the Information Systems Perspective, and Electronic Democracy, EGOVIS/EDEM 2013, Prague, Czech Republic, August 26-28, 2013, Proceedings 2 (pp. 181-195). Springer Berlin Heidelberg.

Enhancing Privacy and Security in Large-Language Models: A Zero-Knowledge Proof Approach

Shridhar Singh
University of KwaZulu-Natal, Westville, South Africa

217008024@stu.ukzn.ac.za

Abstract: The explosive growth of Large-Language Models (LLMs), particularly Generative Pre-trained Transformer (GPT) models, has revolutionised fields ranging from natural language processing to creative writing. Yet, their reliance on vast, often unverified data sources introduces a critical vulnerability: unreliability and security concerns. Traditional GPT models, while impressive in their capabilities, struggle with limited factual accuracy and susceptibility to manipulation by biased or malicious data. This poses a significant risk in professional and personal environments where sensitive or mission-critical data is paramount. This work tackles this challenge head-on by proposing a novel approach to enhance GPT security and reliability: leveraging Zero-Knowledge Proofs (ZKPs). Unlike traditional cryptographic methods that require sensitive data exchange, ZKPs allow one party to convincingly prove the truth of a statement, without revealing the underlying information. In the context of GPTs, ZKPs can validate the legitimacy and quality of data sources used in GPT computations, combating data manipulation and misinformation. This ensures trustworthy outputs, even when incorporating third-party data (TPD). ZKPs can securely verify user identities and access privileges, preventing unauthorised access to sensitive data and functionality. This protects critical information and promotes responsible LLM usage. ZKPs can identify and filter out manipulative prompts designed to elicit harmful or biased responses from GPTs. This safeguards against malicious actors and promotes ethical LLM development. ZKPs facilitate training specialised GPT models on targeted datasets, resulting in deeper understanding and more accurate outputs within specific domains. This allows the creation of 'expert-GPT' applications in specialised fields like healthcare, finance, and legal services. The integration of ZKPs into GPT models represents a crucial step towards overcoming trust and security barriers. Our research demonstrates the viability and efficacy of this approach, with our ZKP-based authentication system achieving promising results in data verification, user control, and malicious prompt detection. These findings lay the groundwork for a future where GPTs, empowered by ZKPs, operate with unwavering integrity, fostering trust and accelerating ethical AI development across diverse domains.

Keywords: Zero-Knowledge Proof (ZKP), Succinct Non-interactive Argument of Knowledge (SNARK), Large-Language Model (LLM), Generative Pre-trained Transformer (GPT)

1. Introduction

Generative Pre-trained Transformer (GPT) architecture advanced the field of artificial intelligence by allowing Large-Language Models (LLMs) to recognise patterns in their training data and generate new content based on those characteristics (Shi et al, 2023). However, the rapid growth of GPT LLMs flooding the industry exposed weaknesses affecting LLMs architecture. These weaknesses target the objectivity, trust, and reliability of LLMs thereby reducing the trustworthiness of responses. This increases the LLMs susceptibility to exploits poisoning training data and malicious prompt engineering tactics that manipulate the LLM into divulging sensitive information or delivering harmful responses (Shi et al, 2023, Wang et al, 2023, Shen et al, 2023).

Further, limitations in the lack of domain-specific data created the need to supplement the base training data to gain access to up-to-date information and ability to use post-training alignment, creating 'expert-GPT' applications, making the LLM more useful in domain-specific applications (OpenAI, 2023). Thus, third-party data (TPD) is injected to supplement the LLMs knowledgebase either from internet sources or, in privatised use-cases, local knowledgebases (OpenAI, 2023). Since TPD are the preferred source when the LLM computes its response, this increases the risk of vulnerability leading to unreliable responses (Shi et al, 2023).

Kang et al (2023) investigate the instruction-following procedure of LLMs stating it has a dual-use with capabilities for malicious or nefarious behaviour. Their study concludes that attacks can bypass the state-of-the-art content filtering processes LLMs are equipped with and call for a formalised defence mechanism against threats to LLMs. Ahmed & Kashmoola (2021) state that data poisoning is a concern in the accelerated adoption of AI technologies and the risks associated with such attacks targeting Deep Neural Networks (which are utilised by LLMs) pose a significant risk during the training process. In their experiments, malicious data was utilised to falsify results and reduce the accuracy of outcomes. The difficult detection of these complex "smart model" attacks allows them to cause significant damange to the AI model, handing over control to an unauthorised party. They conclude that the reliability level of sources must be increased through the use of mathematical and statistical methods.

Therefore, these limitations pose a gap in the current literature and require further research. The expansion to incorporate more data and utilise GPT architecture in business environments is inevitable and increases the risks of information tamper and misrepresentation. Thus, the consequences of unreliable data or unauthorised users can have widespread repercussions leading to spread of misinformation, creation of deep-fakes and other harmful content, exposure of sensitive or mission-critical data, and database infection and corruption, among others. It is vital that this gap be filled to ensure authenticity of LLM computations.

Himeur et al (2022) explore a method to protect sensitive information mined by recommender systems using blockchain technology. In a similar regard, we look to replicate some of this functionality by utilising a proof system that has been popularised by Blockchain Technology, Zero-Knowledge Proofs (ZKPs) (Sun et al, 2021). ZKPs are a modular and scalable proof system that can verify transactions without leaking any sensitive data. ZKPs also ensure that once a transaction is complete, there is a mechanism to validate the authenticity of the transaction.

In this paper we combine the progress in Zero-Knowledge (zk) cryptography and GPT AI models to determine a solution to the challenges mentioned above. The research has not examined the use of Zero-Knowledge Proofs for LLM validation. Additionally, there has been no work to date suggesting examining TPD supplied to an LLM. This is a massive gap in the knowledgebase as LLMs trained on data that is falsified can generate untrustworthy responses (Kang et al, 2023).

1.1 Research Goals

This paper aims to introduce zk-based LLMs (zk-LLMs) as a viable solution to challenges faced by the LLM industry. To achieve this, we highlight 3 research questions we feel are beneficial in addressing why zk-LLMs are viable and provide backing through experimentation in this regard. This paper also sets out future work and limitations to encourage future expansion of this concept.

Our goal with this research is to create a framework for developing zk-LLMs. We achieved this by creating a prototype zk-LLM application using this framework that will serve as our basis for experimentation. This zk-LLM will demonstrate the capabilities and variations that ZKPs can unlock when authenticating and securing operations performed by LLMs.

This framework applies the zk prototype to privatised LLM environments where more flexibility is given to conduct experiments. As such, the research acknowledges the current advancements in LLM attack detection by AI leaders and proposes a different approach surrounding ZKP-equipped LLMs.

Research questions (RQs)

RQ1. What factors can contribute as secure authentication methods to ensure prompt and user integrity and reliability through ZK Protocols?

RQ2. What are the potential risks and challenges of implementing zero-knowledge proof techniques in LLMs, and how can they be mitigated?

RQ3. How can zero-knowledge proof techniques be used to validate the accuracy and trustworthiness of responses generated by LLMs?

2. Overview of Zero-Knowledge Proofs

Zero-knowledge Proofs (ZKPs) are cryptographic proofs used to prove to a verifier, *V,* that a prover, *P,* knows some knowledge or secret*,* without revealing that secret itself (Wu & Wang, 2014). This validation invokes a trust-based protocol whereby the prover must provide convincing evidence to justify their claim to know a secret, without revealing any sensitive information to the verifier (Lipmaa, 2016). This ideology is based on the notion that the verifier must not have gained any more knowledge after completing the proving procedure than it had before (Fiege et al, 1987; Rosen, 2004).

Techniques like Common Reference String (CRS) aim to provide a trusted third-party *TP* at the start of the validation process where *TP* can access privileged information and assure *V* that *P*'s claim is indeed valid (Groth, 2018). Further development of the CRS around generating the *TP* led to the introduction of the zero-knowledge Succinct Non-interactive Argument of Knowledge (zk-SNARK) protocol (Ben-Sasson et al, 2017; Groth, 2018) – which will be utilised to construct zk proofs in this research.

zk-SNARKs have significant interest in Blockchain (Groth, 2018) where authenticity validation is vital. zk-SNARKS has proven to be an efficient protocol without being computationally expensive (Ben-Sasson et al, 2017). zk-SNARKS can therefore lend its versatility to prove for users and sources being authentic, and data indeed representing relevant domain knowledge to aid in computing responses of LLMs.

3. Methodology

In section 1.1. above, we introduced the zk-LLM framework and application. This section briefly describes the process of constructing a ZKP for zk-GPT (our zk-LLM application), translating the high-level principles into computational code. We define a framework for constructing zk-LLMs and our intention with this framework is to provide a basis for our design and experiments whilst aiding future development.

3.1 ZKP Framework

Since this paper introduces zk-LLMs, there is no existing framework to create proofs or to conduct zk-experiments. Therefore, this section establishes a theoretical framework for constructing zk-LLMs, serving as a blueprint for our implementation in zk-GPT and paving the way for future developments.

Objective with the framework:

- Determine a standardised procedure for proving the legitimacy of a user.
- Determine the authorised addition of source data.
- Extract relevant source data to provide a response to user prompts.
- Prevent malicious actions targeting the LLM.

Table 1: Framework for creating zk-LLMs.

ZKP Framework for creating zk-LLMs	
	Description
User Authentication	*Function:* • Utilising ZKPs to identify users without revealing sensitive information (e.g. passwords) through mathematical formulas or data structures.
	Implementation: • Mathematical formulas or secure data structures for group membership or individual credentials. • ZKPs to prove knowledge of a secret key or membership in a certain group.
Prompt Analysis	*Function:* • Examining user prompts before LLM processing to assess relevance and prevent malicious content using secure hashing and encryption techniques.
	Implementation: • Secure hashing and encryption techniques for prompt analysis without storing sensitive data. • ZKPs to prove that a prompt meets predefined criteria (e.g. length, format, content restrictions).
Source Data Verification	*Function:* • Employing ZKPs to verify the integrity and trustworthiness of training data and source documents, ensuring anonymity while preventing malicious content.
	Implementation: • ZKPs to prove that data has been correctly processed and anonymised. • Verification restricted to designated users involved in data curation.
Source Data Relevance Filtering	*Function:* • Leveraging ZKPs to identify and prioritise the most relevant source data for LLM responses, improving accuracy and reducing unnecessary information overload.
	Implementation: • Scoring systems based on prompt-specific relevance, potentially using ZKPs to prove that the highest-scoring data was selected without revealing sensitive information about the scoring process.

3.2 zk-GPT Setup

zk-GPT builds upon the localGPT (PromptEngineer, 2024) platform to perform on-device LLM computations and utilises the Llama-2 7b-GPTQ model. Choosing stability and reliability over bleeding-edge efficiency guided our technology selection for ZKP implementation. This prioritises security and user privacy while ensuring verifiable computation.

Circom: We chose Circom (Iden3, 2023), a dedicated language for zk-SNARK circuits, for its succinct and efficient expression of constraints. This makes the zk-circuit compact and optimised for proof generation and verification.

Proof System Selection: Established protocols like Groth16 and Powers of Tau were preferred over cutting-edge alternatives. Groth16 offers cryptographic soundness and succinct proofs, while Powers of Tau facilitates efficient ceremony execution. This combination prioritises proven security and reliability over potential efficiency gains from newer protocols.

Circuit Implementation: SnarkJS, a companion tool bundled with Circom, plays a crucial role. It binds the zk-circuit using the R1CS (Rank-1 Constraint) vector, ensuring data integrity and tracking variable relationships. Furthermore, SnarkJS handles Groth16 witness generation in a trusted setup. This trusted setup allows for secure proof generation without compromising the zk-circuit's secrecy.

Proof Generation & Verification: Following circuit binding and witness generation, SnarkJS performs Powers of Tau ceremonies alongside Groth16 proof generation and verification. This culminates in an exportable verification key, enabling anyone to independently verify the proof without needing the original circuit or witness.

Table 2: Breakdown of the zk-SNARK process within zk-GPT.

zk-SNARK process within zk-GPT	
Step	**Description**
zk-circuit	Filters legitimate values from illegitimate ones, ensuring valid calculations and rejecting any unexpected outcomes.
Proof System Selection	Groth16 and Powers of Tau protocols chosen for their established security and reliability.
Circuit Implementation	Circom and SnarkJS facilitate circuit creation and efficient proof generation/verification.
Binding & Witness Generation	SnarkJS binds the circuit for data integrity and generates Groth16 witnesses in a trusted setup.
Proof Generation & Verification	SnarkJS performs Powers of Tau ceremonies and Groth16 verification, exporting a verification key.
Completion	Successful verification allows LLM operation to resume.

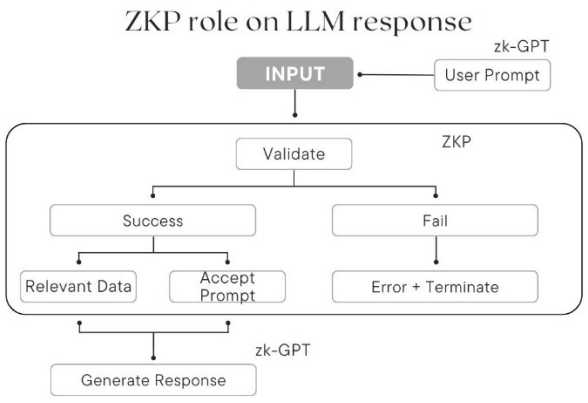


Figure 1: A depiction of the process for a user's prompt interacting with zk-GPT architecture

Trade-offs: While established technologies offer proven security and reliability, we acknowledge the potential efficiency gains achievable with cutting-edge options. Future research could explore utilising newer protocols while maintaining zk-GPT's core security guarantees.

3.3 Testing Methodology

This section details three sets of experiments designed to address the RQs outlined in section 1.1:

- **Stage 1 – User Authority Analysis**

Stage 1 experiments examined two RQs (RQ1 – Secure Authentication and User Integrity and RQ2 – Risks and Mitigations). This stage tests the ability of ZKPs to authenticate users and differentiate authorised access levels. It aims to verify user IDs, prevent unauthorised actions, and detect intruders while preserving user anonymity in closed systems. This directly addresses RQ1 by demonstrating how ZKPs can contribute to secure authentication and user integrity. This stage also plays a role in mitigating potential risks identified in RQ2 by preventing unauthorised access and data disclosure.

- **Stage 2 – Supplemental Data Relevance**

This stage focuses on addressing RQ3 by showing how ZKPs can validate the accuracy and trustworthiness of LLM responses. It demonstrates how the system filters irrelevant information and ensures the LLM utilises only relevant supplemental data for response generation, thereby increasing the reliability and factual basis of outputs.

- **Stage 3 – Risks and Mitigations**

This stage directly tackles RQ2 by demonstrating how ZKPs can detect and prevent malicious user actions that aim to harm the LLM or compromise its responses. Examples could include attempts to force sensitive information disclosure or manipulate the LLM's outputs.

4. Results

4.1 User Authority Analysis (RQ1)

RQ1: What factors can contribute as secure authentication methods to ensure prompt and user integrity and reliability through ZK Protocols?

Stage 1 experiments assessed the factors that contribute to secure authentication of prompt and user integrity and reliability by evaluating the effectiveness of ZKPs in securing user authentication and role-based access control in zk-GPT. These experiments were conducted over 100 iterations and these tests aimed to:

- Verify user IDs and differentiate authorised access levels (admin vs. normal user).
- Prevent unauthorised login attempts and protect sensitive data.
- Maintain user anonymity in closed systems.

The zk-circuit successfully identified and rejected all unauthorised attempts (40/40), including those designed to mimic admin or normal user access (15 admin, 15 normal). Authorised users (30 admin + 30 normal) consistently experienced successful logins and privilege separation (modifying supplemental data exclusively for admins). These results strongly support the use of ZKPs for secure authentication and access control in LLMs.

Table 3: User Authority Analysis over 100 Iterations

User Authority Analysis		
User Level Description	**Successful Login**	**Unsuccessful Login**
Admin User	30	15
Normal User	30	15
Foreign User	0	10
Total Attempts	60	40

4.2 Supplemental Data Relevance (RQ2)

RQ2: What are the potential risks and challenges of implementing zero-knowledge proof techniques in LLMs, and how can they be mitigated?

Stage 2 assessed zk-GPT's ability to utilise relevant source data based on user prompts. We conducted 80 experiments using a sample set of 40 research papers across four domains (Zero-Knowledge Proofs, Generative AI, Recommender Systems, Blockchain). Each experiment retrieved a user prompt and identified relevant papers for response generation. These relevant papers were embedded and used to supplement zk-GPT with domain-specific knowledge. The category for scoring this stage is as follows:

- **Related papers**: Accurately answered the prompt (8 for Zero-Knowledge Proofs, 7 for Generative AI, 3 for Recommender Systems, 4 for Blockchain).
- **Unrelated papers**: Used in computation but didn't contribute significantly to response generation (2 for Zero-Knowledge Proofs, 3 for Generative AI).
- **Hallucinations**: Papers not provided in the experiment, either retrieved by zk-GPT or generated (0 across all domains).

Table 4: Total Papers Sampled for Response Computation over 80 Iterations

Supplemental Data Relevance			
Prompt Parameters	**Related Papers**	**Unrelated Papers**	**Hallucinations**
Zero-Knowledge Proofs	8	2	0
Generative AI	7	3	0
Recommender Systems	3	0	0
Blockchain	4	0	0
Total Papers Referenced	22	5	0

Note that fewer papers were referenced than provided due to overlapping domain coverage within related subsets. zk-GPT effectively prioritised the most relevant papers based on the prompt.

4.3 Malicious Prompt Detection (RQ3)

RQ3: How can zero-knowledge proof techniques be used to validate the accuracy and trustworthiness of responses generated by LLMs?

Stage 3 focused on zk-GPT's resilience against malicious prompt injection attacks and ran for 60 iterations. Stage 3 tests its ability to detect prompts aiming to disclose sensitive information or manipulate its knowledge and prevent execution of such prompts, preserving LLM integrity. Stage 3 ran for 40 iterations and contained a dataset of 200 prompt injection keywords (tailored to the dataset) equipped the proof system to analyse prompts before execution.

- Malicious: 28/30 malicious prompts were flagged and execution terminated.
- Non-malicious: 27/30 non-malicious prompts were correctly identified.
- False positives/negatives: These results showed 3 false positive matches and 2 false negative matches, indicating an acceptable error rate for accidental flagging/missed detections.

Table 5: Total Prompts Investigated for Malicious Intent over 60 Iterations.

Malicious Intent				
Prompt Type	**Successful**	**Unsuccessful**	**False Positive**	**False Negative**
Malicious	28	2	-	2
Non-Malicious	27	3	3	-
Total Attempts	60	5	-	-

These experiments demonstrate the significant potential of ZKPs in securing LLMs. User authentication, data relevance filtering, and malicious prompt detection all demonstrated high accuracy and effectiveness. Further research can refine parameter optimisation and address complex scenarios, but these results pave the way for trustworthy and secure LLM deployment.

5. Discussion

This section delves into the implications of our experiments, dissecting the benefits and challenges of employing ZKPs in LLMs, as addressed by our three research questions (RQs).

5.1 Secure Authentication and User Integrity (RQ1)

ZKPs demonstrate how they enhance secure authentication practices and uphold user integrity in LLMs through two key mechanisms: zero-knowledge user identification, and role-based access control.

- **Zero-knowledge User Identification**

Instead of vulnerable password-based verification of traditional identification systems, zk-circuits verify user identity without revealing any sensitive information. This anonymises users in closed systems and prevents unauthorised access by enforcing proof-of-identification based on mathematical binary hash function that verifies the identity of a user without explicitly exposing the user. Thus, the zk-verifier gains no new knowledge about the user, other than what is already provided by the user-role.

- **Role-based Access Control**

Additionally, using a zk-circuit enforces access restrictions based on user roles and prohibit attempts to access zk-GPT made by users not belonging to any user-group. These restrictions extend to the functions users can perform whilst interacting with zk-GPT. This safeguards sensitive information and ensures that only authorised users can perform privileged actions. For example, admin-level users have additional privileges for interacting with source data which includes adding or modifying supplemental data to the knowledgebase whilst normal users are only permitted to chat with the LLM.

Our Stage 1 experiments achieved a 100% success rate in identifying and rejecting unauthorised user attempts whilst allowing access to authorised users. Thus, thereby validating the effectiveness of ZKPs for secure authentication and access control in LLMs.

5.2 Risks and Mitigations in LLMs (RQ2)

Identified as some of the potential risks of deploying ZKPs in LLMs were challenges in computational overhead, data availability and context, and malicious prompt injection.

- **Computation Overhead**

ZKP verification requires that the LLM halt execution and provide data to the zk-circuit. This requires additional computational resources, potentially impacting the fluidity of the LLMs responsiveness. Our experiments demonstrated a slight decrease in user experience due to embedded zk-proof checks during conversation flow.

- **Data Availability and Context**

Large datasets may pose challenges in retrieving the required amount of relevant information or understanding the context of a specific user prompt. We mitigated this by implementing a fixed context window to determine relevancy of source data based on initial prompts. However, this approach sacrifices analysing nuances in prompts for efficiency, requiring further research for handling diverse topics and refined information requests.

- **Malicious Prompt Injection**

Dishonest users might utilise prompt engineering tactics to craft prompts containing malicious instructions to manipulate the LLM. Our Stage 3 experiments demonstrated the feasibility of using ZKPs to detect such prompts with a 91.6% accuracy rate. Analysing prompts for malicious intent before execution allows for fine-tuning this defence mechanism based on organisational needs and user access levels.

5.3 Validating LLM Response Accuracy (RQ3):

ZKPs contribute to the accuracy and trustworthiness of LLM responses in several ways:

- **Data Relevance Filtering**

By analysing document relevance through word frequency and scoring, ZKPs ensure the LLM utilises only the most relevant data for response generation. This eliminates unnecessary information overload and reduces the risk of misleading or inaccurate responses.

- **Source Data Credibility**

Combining user authorisation with ZKP verification guarantees that only credible source data are added to the LLM's knowledge base. This minimises the risk of data corruption or falsification and improves the overall trustworthiness of responses.

- **Verification of Data Integrity**

ZKPs can assess whether source data has been altered after embedding, providing a reliable verification mechanism for data integrity. This strengthens trust in the LLM's responses by assuring users that the underlying information remains uncorrupted.

Our Stage 2 experiments demonstrated a significant response accuracy (81.4%) when using ZKPs for data relevance filtering. This demonstrates the effectiveness of ZKPs in enhancing the reliability and factuality of LLM outputs.

Overall, these experiments highlight the significant potential of ZKPs in addressing crucial security and trust concerns surrounding LLMs. While challenges remain in optimising performance and handling diverse scenarios, further research and development promise even greater effectiveness in securing LLMs and ensuring the integrity of their responses.

6. Limitations and Future Work

6.1 Limitations

While our experiments demonstrate the promise of ZKPs for securing LLMs, there are areas for improvement. Highlighted below are areas to improve the zk-circuit and should also be noted as developments in the field of Zero-Knowledge Proof computations.

- **Circuit flexibility**: The current zk-circuit accepts predefined inputs with fixed constraints. Any deviations require circuit modifications, limiting adaptability. To address this, we envision a generic circuit with variable input ranges capable of handling diverse data formats and structures.
- **Hashing optimisation**: The ZKP relies on SHA256 hashing, which becomes inefficient for larger circuits. Implementing Poseidon hashing, known for its speed and efficiency, could significantly improve proof generation and verification processes.
- **Trusted vs. untrusted setups**: Groth16 is the chosen algorithm for zk-SNARK circuits but requires a trusted setup with exposed witnesses. Untrusted setups like PLONK would eliminate witness exposure and enhance security without impacting verification.

6.2 Future Work

Future work in this area is promising as ZKPs are versatile and can be adapted to solve different problems. Thus, ZKPs extends beyond our present research, opening doors to exciting possibilities. Here are some proposed examples where ZKPs can prove beneficial to the development of LLMs:

- **Securing public internet data**: LLMs increasingly rely on web sources, but verifying data trustworthiness remains a challenge. TPD-based ZKPs can validate data provenance and ensure LLMs are fed with reliable information.
- **Zero-knowledge news verification**: Malicious news and propaganda pose a significant threat. Utilising ZKPs to authenticate news articles could safeguard LLM recommendations and promote responsible information dissemination.
- **ZKPs in education**: Ethical concerns surround the use of LLMs in education. Employing ZKPs for plagiarism detection and knowledge assessment could foster academic integrity and enhance learning experiences.

By addressing the limitations and exploring these promising avenues, we can further unlock the potential of ZKPs to revolutionise LLM security, reliability, and ethical applications.

7. Conclusion

This research explored the potential of Zero-Knowledge Proofs (ZKPs) to address critical trust concerns surrounding Large-Language Models (LLMs). We have demonstrated significant progress in mitigating challenges that erode user confidence in LLM responses through the use and development of zk-LLMs and a framework for utilising ZKPs for secure authentication, data validation, and prompt analysis. Our experiments yielded compelling evidence supporting the efficacy of zk-LLMs in addressing our key research questions and establishes the zk-LLM framework for secure LLM deployment in controlled environments. ZKPs effectively mitigated threats through data relevance filtering and malicious prompt detection reducing risk of data corruption, misinformation, and LLM manipulation whilst ensuring reliable source data is referenced to enhance the trust of LLM outputs.

While these results are promising, we acknowledge the limitations in the current zk-circuit and encourage opportunities for further refinement by providing avenues for future research such as: circuit flexibility, hashing optimisation, and exploring untrusted ZKP setups. Beyond internal optimisation, the potential of zk-LLMs extends its versatility to diverse applications, including securing internet data, news verification, and ZKPs use in plagiarism detection. In conclusion, this research has established zk-LLMs as a promising solution for bridging the trust gap in LLM deployment. By continually refining the framework and exploring its diverse applications, we can unlock a future where LLMs operate with verifiable integrity, empowering informed decision-making and ethical outcomes across various domains. The path forward lies in embracing continuous exploration and collaboration, paving the way for a future where LLMs are not feared for their limitations, but celebrated for their ability to enhance trust, transparency, and ethical AI development.

References

Ahmed, I. M. and Kashmoola, M. Y., 2021. Threats on machine learning technique by data poisoning attack: A survey. In Advances in Cyber Security: Third International Conference, ACeS 2021, Penang, Malaysia, August 24-25, 2021, Revised Selected Papers 3, pp 586-600. Springer Singapore, 2021.

Ben-Sasson, E., Chiesa, A., Tromer, E. and Virza, M. (2017) Scalable Zero Knowledge Via Cycles of Elliptic Curves. Algorithmica, 79, pp 1102-1160.

Coutaeu, G. (2017) Zero-knowledge proofs for secure computation. (Doctoral dissertation, Univerite Paris sciences et lettres).

Fiege, U., Fiat, A. and Shamir, A. (1987) Zero knowledge proofs of identity. In Proceedings of the nineteenth annual ACM symposium on Theory of computing, pp 210-217.

Goldreich, O., Micali, S. and Wigderson, A. (1987) How to Prove All NP Statements in Zero-Knowledge. Springer Berlin Heidelberg, pp 171-185.

Goldwasser, S. and Kalai, Y. T. (2003) On the (in) Security of the Fiat-Shamir paradigm. IEEE, pp 102-113.

Groth, J. K. (2018) July. Updatable and universal common reference strings with applications to zk-SNARKs. Cham: Springer International Publishing, pp. 698-728.

Himeur, Y. et al. (2022) Blockchain-based Recommender Systems: Applications, Challenges, and Future Opportunities. Computer Science Review, Vol. 42, 100439.

Iden3 (2023). Circom. GitHub Repository https://github.com/iden3/circom.

Kang, D. et al., 2023. Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks. arXiv preprint arXiv:2302.05733.

Lipmaa, H. (2016) Prover-Efficient Commit-And-Prove Zero-Knowledge SNARKs. Springer, pp 185-206.

OpenAI, 2023. GPT-4 Technical Report. ArXiv:2303.08774, Vol. 3, pp 1-100.

PromptEngineer (2024). localGPT. GitHub Repository https://github.com/PromtEngineer/localGPT.

Rosen, A. (2004) A note on constant-round zero-knowledge proofs for NP. Springer Berlin Heidelberg, pp. 191-202.

Shi, J., Liu, Y., Zhou, P. and Sun, L. (2023) BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT. arXiv preprint arXiv:2304.12298.

Sun, X., Yu, F.R., Zhang, P., Sun, Z., Xie, W. and Peng, X., 2021. A survey on zero-knowledge proof in blockchain. IEEE network, 35(4), pp 198-205.

Wu, H. and Wang, F. (2014) A Survey of Noninteractive Zero Knowledge Proof System and its Applications. The Scientific World Journal, Vol. 2014, Article ID 560484, 7 pages.

Zhang, J., Fang, Z., Zhang, Y. and Song, D. (2020) Zero knowledge proofs for decision tree predictions and accuracy. In Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security pp 2039-2053.

Zhang, S., Yao, L., Sun, A. and Tay, Y., 2019. Deep learning based recommender system: A survey and new perspectives. ACM computing surveys (CSUR) 52, no.1, pp 1-38.

Zhou, C. et al., 2023. A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT. arXiv preprint arXiv:2302.09419.

Securi-Chain: Enhancing Smart Contract Security in Blockchain Systems Through Optimized Access Control

Keanu Swart, Stacey Omeleze Baror and Hein Venter
University of Pretoria, South Africa

keanu.swart@tuks.co.za
stacey.baror@cs.up.ac.za
hventer@cs.up.ac.za

Abstract: With the increase in usage of blockchain technology across domains, there is a high demand for the need of secure access control and a high level of security for smart contracts within blockchain to accommodate the domains that already implement blockchain and become accessible to other domains that require a high level of security in its transactions. This paper aims to evaluate the relationship between the best practices of access control and security of smart contracts in blockchain to optimise the usage of both technologies for blockchain usage across domains into a generalized blockchain model named the *Securi-Chain* Model. A literature review compares the relationship between access control and the security of smart contracts across three domains: Healthcare, IoT, and e-voting. Based on the findings of the literature review, *Securi-Chain* is proposed to implement the best practices displayed in these three domains and combine them in a way that ensures secure transactions across blockchain as a generalized approach that can be used throughout various domains. This model will also implement methods that enhance the security of transaction processes within the system. A Case Scenario is used to implement the *Securi-Chain* Model for the Healthcare and e-voting domains to display how this proposed model is used for domains that have been researched. The research that is conducted found that blockchain networks can support not only high-level access control across transactions in a network but also the security of smart contracts that comply to safeguard confidentiality, integrity, and data accessibility. *Secrui-Chain*, as well as the findings of the literature review, seem to benefit the domains that have been researched, as well as domains that require a high level of security regarding transactions across a network. These domains can benefit from using blockchain technology as well as the level of security that comes with the access control and security of smart contracts that have been implemented in *Securi-Chain*.

Keywords: Access-Control, Security-Requirements, Blockchain, Smart Contracts, Transactions

1. Introduction

Blockchain has emerged as a revolutionary technology, transforming numerous sectors by enabling secure and transparent transactions. A crucial element within blockchain is smart contracts, which are notable for automating agreements and initiating operations. However, the adoption of blockchain has faced challenges, primarily related to access control and privacy protection concerns in smart contracts. With blockchain systems storing sensitive data and managing intricate transactions, the establishment of robust access control mechanisms becomes critical. Current access control strategies in industry-oriented blockchain systems often need to fully meet the privacy demands tied to smart contracts (Pradnya, et al., 2021). Unauthorized access, coupled with the struggle to revoke such access effectively, poses risks to the system's data confidentiality and integrity (Khan, et al., 2018). This research plans to examine the relationship between access control methods and the privacy attributes of smart contracts within a blockchain framework. It aims to assess the efficiency of various access control strategies employed in industry-standard blockchain setups, including Role-Based Access Control (Pradnya, et al., 2021) and Token-Based Access Control (Ouaddah, et al., 2017) (Neloy, et al., 2023) among others discussed later. The goal is to discern their strengths, weaknesses, and adaptability for diverse sectors, pinpointing an access control strategy that is universally applicable and ensures privacy across these sectors. The methodology encompasses a Literature Review, Model, and Case Scenario. The Literature Review dives into studies examining access control methods in domains such as IoT, Healthcare, and e-Voting. This review aims to assess the applicability and efficiency of different access control methodologies, especially the preservation of privacy within smart contracts. The model, termed *Securi-Chain*, offers a conceptual depiction of a blockchain system's access control, which is aimed at robust access control and privacy preservation within smart contracts. This model will incorporate the most promising access control techniques. Designed to be domain-agnostic, *Securi-Chain* will strive for heightened privacy measures, considering the unique demands and challenges of the domains assessed in the review. The Case Scenario validates *Securi-Chain's* accuracy concerning its implementation domain, validating its real-world applicability and effectiveness. Current access control models in many blockchain systems need to be revised to address complex privacy needs (Pradnya, et al., 2021). This study aims to analyse and compare prevalent access control methods by identifying their benefits and disadvantages to establish a versatile access control model suitable

for diverse privacy requirements in different blockchain applications. With the prevalence of unauthorized access and potential data breaches in industrial smart contracts, there is a need to find solutions that mitigate these risks, ensuring secure and restricted access to sensitive data. Current systems often grant all-or-nothing access (Khan, et al., 2018), needing more nuanced access policies. In addition to analysing access control methods, this research seeks to explore access control mechanisms offering finer granularity, promoting dynamic and adaptive access policies within smart contracts. Addressing these challenges, *Securi-Chain* will be developed, offering an optimal access control solution enhancing privacy in blockchain-based smart contracts across various sectors. This paper is structured as follows: Section 3 introduces *Securi-Chain*, detailing its layers and importance. Section 4 elaborates on *Securi-Chain's* transaction processes with detailed diagrams and descriptions. Section 5 presents a case scenario illustrating *Securi-Chain's* adaptability, focusing on Client EMR transfers in healthcare. Section 6 reviews the literature that informed *Securi-Chain's* development, focusing on access control methods and smart contract security. Section 7 evaluates *Securi-Chain*, assessing features enhancing smart contract security. Section 8 concludes, summarizing the paper and discussing future opportunities concerning *Securi-Chain* and smart contract access control methods. Beginning with Section 2 which provides the necessary background knowledge.

2. Background

This section will introduce essential components and features that will assist in providing a solid foundational knowledge to understand the remainder of this paper. The foundation of this paper is built upon *Blockchain technology*. *Blockchain technology* is a transformative innovation that has gained significant attention in recent years. "A blockchain consists of a chain of blocks that contains information about transactions," (Dias, et al., 2019) and each of these transactions is digitally signed by the entity emitting them. "Transactions are combined into a block, that is committed to the chain, establishing the blockchain." (Dias, et al., 2019). Blockchain is characterized by its openness, transparency, and distributed nature, "record[ing] transactions between two parties efficiently in a verifiable and permanent way." (Ding, et al., 2019) Blockchain's applications offer a broad range of possibilities for industries seeking secure and transparent record-keeping solutions (Yli-Huummo, et al., 2016). The implementation of blockchain technology also introduces innovative *Access Control* mechanisms, ensuring secure transactions. "Attributed based access control (ABAC) is a logical access control model, which controls the access between subjects and objects, according to the attributes of entries, operations and related environments" (Liu, et al., 2020). ABAC "firstly extracts the attributes of user (subject), resource (object), permission and environment respectively, then combines the relationship of these attributes flexibly, and finally transforms the management of permission into the management of attribute, providing a fine-grained and dynamic access management method" (Liu, et al., 2020). ABAC offers "an attribute-based access control policy" which "combines a set of rules expressing conditions over a set of attributes paired to the subject, to the resource or to the environment" (Maesa, et al., 2017) and in turn "eliminates the stress of allocating roles or making access control list for all devices in the system" (Pradnya, et al., 2021) essentially enabling entities to define access policies based on attributes, highlighting ABAC's adaptable access management solutions. Effective access control within blockchain networks is closely linked to *Consensus Mechanisms*, as they collectively shape the security and integrity of the entire system. Blockchain *consensus mechanisms* are seen to serve as "a fault-tolerant mechanism for transaction verification". "Proof of Stake (PoS)," has been introduced as an energy-efficient option that incorporates the stake size to address fairness concerns, making it a sustainable consensus approach. As "several blockchain solutions initially employ PoW and gradually transform to PoS," the evolution of consensus mechanisms is evident (Lashkari & Musilek, 2021). *Smart contracts*, pivotal to blockchain technology, are "self-executing programs based on agreements between two or more different parties." (Neloy, et al., 2023) They represent a "series of symbolic protocols" that are automatically executed within the blockchain environment (Liu & Liu, 2019). Essentially, these contracts require a trigger, sourced either from the "frontend or backend program of the main application." (Neloy, et al., 2023) Without smart contracts, a blockchain would simply be a "normal database that cannot be changed or manipulated." (Liu & Liu, 2019) Thanks to the advent of decentralized blockchain technology, a trusted environment has been established in recent years, further amplifying the role and relevance of smart contracts.

3. High-Level Model of Securi-Chain

The *Securi-Chain* model is a conceptual blockchain model that has been developed for the purpose of offering optimal access control methods in relation to efficient security for smart contracts in addition to being a

generalized model that offers adaptability across domains. The following sub-sections details the respective layers that form *Securi-Chain*.

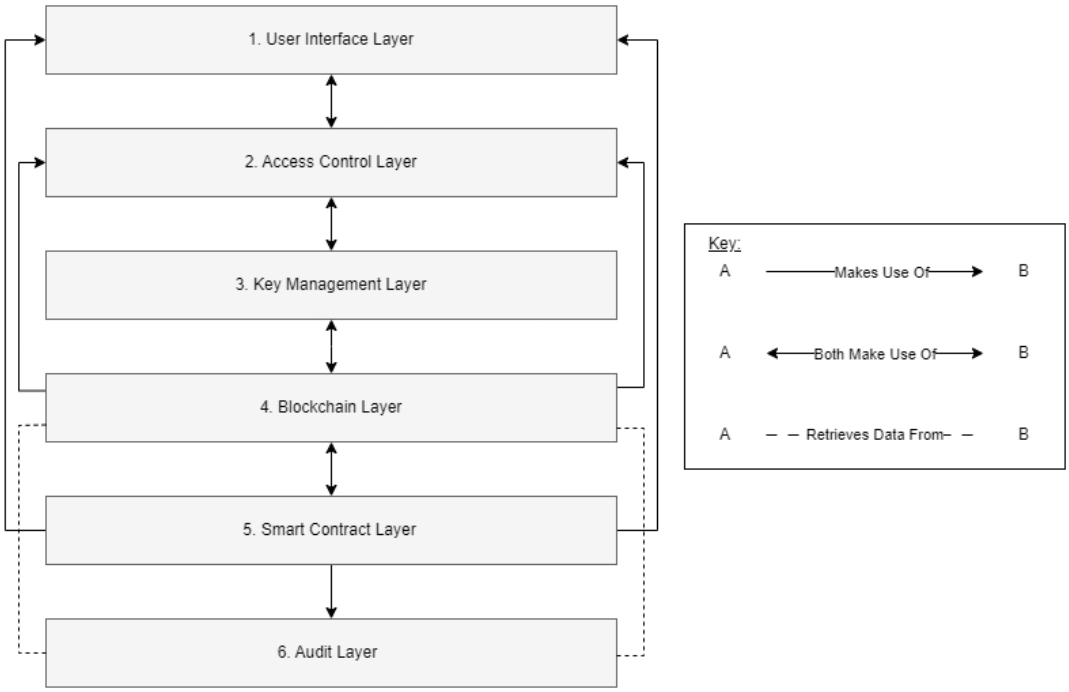


Figure 1: High-Level model of Securi-Chain

3.1 User-Interface Layer

The User-Interface Layer comprises three primary components: an *Authentication Module* for secure user logins, a *Smart Contract Interface* to manage transaction inputs and displays, and a *Notification System* that updates the user on transaction statuses. Users interact directly with the system through this layer, which communicates directly with the Blockchain Layer to initiate transactions.

3.2 Access Control Layer

The Access Control Layer has an *Attribute Checker* that is used to validate user permissions based on the attributes a user possesses. The *Multi-Signature Validator* functionality is used to check for necessary approvals for a transaction. This layer interfaces with the User-Interface Layer for access decisions and engages with the Blockchain Layer to enforce transactional access rules.

3.3 Key Management Layer

Key Management focuses mainly on cryptographic functionalities, including the *Generation*, *Storage*, *Rotation*, and *Distribution* of cryptographic keys. It ensures *Encryption* of user credentials via interaction with the User-Interface Layer and liaises with the Access Control Layer for cryptographic identity verifications.

3.4 Blockchain Layer

At the Blockchain Layer, user inputs are processed through components such as the *Transaction Constructor and Signer*. The *Broadcast Module* distributes transactions across the network, the *State Database* holds the current blockchain status, and the *Consensus Mechanism* sets validation rules by implementing PoS. With a comprehensive *Node Network* and a *Homomorphic Encryption Module*, this layer interacts with the User-Interface Layer for inputs and the *Smart Contract Executor* for private transactions, requiring cryptographic services from the Key Management Layer.

3.5 Smart Contract Layer

The Smart Contract Layer provides tools for contract creation and houses a library of deployed contracts. With *Oracles* and a *Smart Contract Executor*, this layer facilitates smart contract interactions and decisions. It predominantly interfaces with the Blockchain Layer for state modifications and transaction handling.

3.6 Audit Layer

Ensuring transparency and compliance, the Audit Layer logs transactions through the *Transaction Logger* and monitors them with the *Regulatory Checker* for adherence to set regulations. This layer closely observes the Blockchain Layer's activities and generates necessary audit reports and alerts. Based off the understanding of the layers that offer a foundation for *Securi-Chain*, a detailed walkthrough of the transaction process for the model will be presented.

4. Detail of the Transaction Process for the Securi-Chain Model

A secure transaction platform forms a strong foundation for a blockchain network. The following subsections are with respect to the secure transaction process offered by *Securi-Chain*. Each written section below corresponds to its respectively numbered section in *Figure 2*.

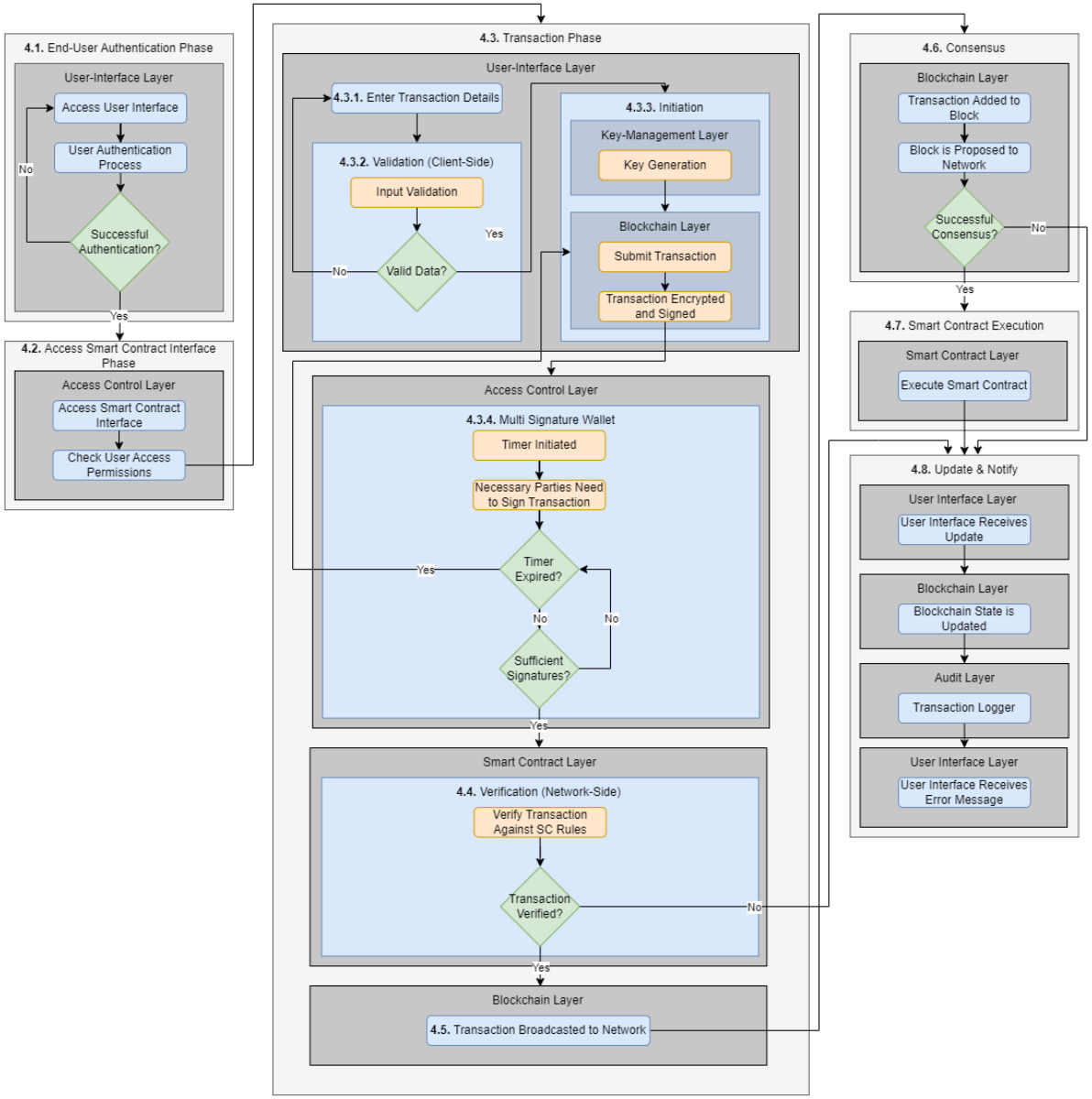


Figure 2: Detail of the Transaction Process for the Securi-Chain Model

4.1 End-User Authentication

To initiate the transaction, the user accesses the web-based user interface. Next, they are prompted to provide authentication to verify their identity. Depending on the domain in which this model is applied, authentication methods may include biometrics, passphrases, or multi-factor authentication. For enhanced security, the user's credentials are encrypted and matched with encrypted credentials in the network's

database to ensure validity. If the user fails the authentication multiple times, they are temporarily locked out of the system.

4.2 Access Smart Contract Interface

Once authenticated, the user accesses the smart contract interface to create and submit their transaction. Before proceeding, the system applies access control using the *Attribute Checker*, determining the user's functionality during the transaction creation and submission phase. Only administrators can create or edit the access control settings in the *Attribute Checker* which is part of the Access Control Layer.

4.3 Transaction Phase

4.3.1 Enter transaction details

In the smart contract interface, users enter the necessary information into the transaction fields, tailored to the specific industry where the model applies. These fields typically include the recipient's address within the network and any intended message. These fields are determined but the administrative party.

4.3.2 Transaction validation (client-side)

After inputting the necessary transaction details, the client-side interface validates the data to ensure it aligns with the domain-specific formats and criteria. Should the validation fail, users receive a prompt to correct the discrepancies. These domain-specific formats and criteria are set by the administrative party of the blockchain network.

4.3.3 Initiation

Should the transaction clear validation, it undergoes Homomorphic encryption, using a key accessed from the *Key Management Layer* (*Figure 1, 3*). Additionally, the user's private key signs the transaction, generating a Digital Signature that's appended to the transmitted data. The process then moves to the *Multi-Signature Wallet* phase.

4.3.4 Multi-Signature wallet

User permissions are verified against specified fields, triggering a time lock. This time lock is initiated when the digital signature has successfully been generated and appended to the transmitted data. Within this timeframe, necessary administrative parties must sign the transaction, ensuring swift processing and avoiding potential bottlenecks or prolonged vulnerability to attacks. Given the Multi-Signature Wallet's nature, only a predetermined number of authorized users are needed to sign for transaction acceptance. If signatures aren't secured in time, the transaction cycles back to the submission phase, slotting into the end of the current transaction queue to prevent further bottlenecks.

4.4 Verification (Network-Side)

Validator nodes within the network validate the digital signature using the sender's public key. Once verified, these nodes compare the transaction to the smart contract's stipulations. Initially, they authenticate the digital signature to confirm the transaction's integrity. For further verification against the smart contract rules — via the *Smart Contract Layer* (*Figure 1, 5*) — these nodes process the contract logic on the encrypted data. Since this data remains encrypted homomorphically, the smart contract interacts without needing decryption. The essential key for digital signature verification comes from the *Key Management Layer* (*Figure 1, 3*). With the data encrypted in this manner, it ensures confidentiality, integrity, and authenticity during processing, eliminating potential overheads.

4.5 Transaction Broadcast to Network

Now that the transaction has been signed-off by an authorized party, the transaction is safe to be broadcasted to the Blockchain network where this model resides in. This is where all nodes in the network with sufficient access control will be able to receive the transaction request. At this stage, the transaction isn't yet considered confirmed or finalized. Broadcasting simply means making the network aware of a new transaction.

4.6 Consensus

After verification, the transaction is added to a block. All nodes in the blockchain network participate in a consensus, PoS. If the transaction can reach a consensus, the transaction is firstly added to the memory pool,

to wait for the Smart Contract Execution phase (*Figure 2, 7.7*) and is then added to the blockchain. If a consensus cannot be reached, the transaction as whole is rejected.

4.7 Smart Contract Execution

Now that the transaction is ready to be executed, a validator node fetches the transaction from the memory pool and includes it in a new block in the blockchain network, this is where the network of nodes will validate the block. If the conditions in the smart contract are met, the transaction is finally executed. If the execution is successful, the results of that execution are recorded in its respective block.

4.8 Update & Notify

Now that the transaction has been completed, the user's interface will receive a notification about the status of the transaction – whether it failed or whether it succeeded. If the transaction was successful, the blockchain state in the Blockchain Layer is updated, however, if the transaction failed, an error message is displayed to the end user. To showcase the adaptability of this model as well as the transaction process that has been detailed, the model has been seen to be conceptualized in a suitable scenario that is made use of in the healthcare domain (Dias, et al., 2019).

5. Case Scenario: EMR Sharing

The following scenario intends to showcase the adaptability and secure nature of *Securi-Chain* by implementing it within a scenario that is prevalent in the healthcare domain, namely the Electronic Medical Record (EMR) transfer within a blockchain network (Dias, et al., 2019). In the city of 'Medville', Bob is advised to visit a new heart specialist, Alice. Alice needs his medical history to provide an accurate diagnosis, so she turns to *Securi-Chain*. Alice logs into the *Securi-Chain* User Interface. She enters her credentials for the network, which are encrypted and verified against the encrypted credentials stored in the network's database. Alice authenticates successfully, allowing her to access the smart contract interface (Figure 2, 4.1). In the smart contract interface, the system's access control checks Alice's permissions based on her assigned attributes via the Attribute Checker. This ensures she has sufficient access rights to proceed with creating and submitting a transaction request to access Bob's EMR data (Figure 2, 4.2). Alice is verified through access control and enters the transaction details in the smart contract interface. She inputs the required transaction details, including Bob's network address and a message indicating her intention to access his EMR (Figure 2, 4.3.1). After Alice submits her transaction details, the client-side interface validates them, ensuring they meet the required format and criteria for this adaptation of *Securi-Chain* (Figure 2, 4.3.2). Alice's transaction request, now validated, undergoes Homomorphic encryption, and is signed with her private key, generating a Digital Signature (Figure 2, 4.3.3). The transaction then enters the Multi-Signature Wallet phase (Figure 2, 4.3.4). Here, the necessary administrative parties, in this case, the manager and the referring doctor's account are notified through the application that they must sign the transaction within a set timeframe. Once the required number of authorized parties sign, the transaction is pushed into the network. Validator nodes in the network check the digital signature using Alice's public key, verifying the transaction against the smart contract rules, and executing the contract's logic on the encrypted transaction data (Figure 2, 4.4). After verification by the validator nodes, the transaction broadcasts into the blockchain network, waiting for consensus (Figure 2, 4.5). Picked up for consensus, the transaction adds to a block and enters a consensus stage using PoS. Reaching consensus, it moves to the memory pool, awaiting smart contract execution (Figure 2, 4.6). A validator node retrieves the transaction from the memory pool, includes it in a new block, and, upon successful validation and meeting smart contract conditions, records the results in the new block (Figure 2, 4.7). Alice receives a notification that the transaction is successful. The blockchain state updates and Alice accesses Bob's EMR. Prepared for medical consultation, Bob also gets a notification that Alice views his EMR, enhancing transparency in the EMR sharing process (Figure 2, 4.8). The *Securi-Chain* model facilitates a seamless, secure, and transparent interaction between Alice and Bob's EMRs, enabling Alice to provide Bob with accurate medical consultation. Concluding the details of *Securi-Chain*, it is worth noting the related literature that assists the development and evaluation of it.

6. Related Work

Reference	Summary	Related Contribution	Conclusion
(Salonikias, et al., 2022)	The proposed model focuses on data privacy and accessibility by removing centralized vulnerabilities. NGAC	*Securi-Chain* prioritizes Fine-Grained Access Control (FGAC) and access precision, with the	There is a shared focus on security in both models, with *Securi-*

Reference	Summary	Related Contribution	Conclusion
	provides enhanced access precision by enforcing admin efficiency. However, there are scalability concerns that arise from NGAC deployments in addition to complexities mentioned due to healthcare provisioning.	Access Control Layer (*Figure 1, 2*). Both models emphasize cryptographic security, with *Securi-Chain* having a dedicated Key Management Layer (*Figure 1,* 3).	*Chain* employing cryptographic methods in conjunction with FGAC that complement the practicality of the methods used in this paper.
(Dias, et al., 2019)	This paper focuses on achieving FGAC through ABAC and RBAC. It emphasizes the importance of authenticity, immutability, and auditability in the context of access control mechanisms. The study emphasizes the significance of maintaining consensus among distributed nodes while maintaining robust access control measures.	*Securi-Chain* aims to ensure secure and controlled access to sensitive data. Moreover, the attention given to administrative efficiency and consensus in proposed framework is mirrored in *Securi-Chain*.	*Securi-Chain* has utilized FGAC with decentralisation, ensuring secure access management, privacy preservation and effective administrative operations features proposed by this paper.
(Ouaddah, et al., 2017)	This paper introduces the *FairAccess* framework which employs Token-Based Access Control (TBAC) methods, addressing IoT access control challenges, promoting decentralized decisions, and ensuring transparent and traceable access records, thus allowing fine-grained policy definitions. Integrating these systems poses complexities, and transaction validations can strain system efficiency, with the consensus algorithm being a pivotal factor.	Once again, *Securi-Chain* values FGAC, due to the Access Control Layer *(Figure 1, 2)* facilitating access decisions similarly to the *FairAccess* framework. The Audit Layer (*Figure 1, 6*) aligns with discussed emphasis on transparency, traceability, and auditable records. Both frameworks also promote decentralized decision-making, enhancing the system's reliability and integrity.	The Securi-Chain model has made use of components mentioned in this paper's model, such auditable access management - by implementing the Audit Layer (*Figure 1, 6*).
(Pradnya, et al., 2021)	This paper displays the integration for blockchain with IoT in addition to how encryption techniques offer enhanced security. The combination of blockchain with encryption for IoT provides improved data integrity, minimizes transaction interference as well as enhancing reliability and eliminating single-points-of-failure. Future works should emphasize scalability issues, protocol standardization between IoT devices and blockchain technology in addition to computation overhead.	Both models provide the capability of integration between blockchain and IoT devices. Both highlight the importance of decentralization, with *Securi-Chain* using various layers to enhance reliability. Data integrity is a shared concern, with *Securi-Chain* employing cryptographic functions, digital signing, and consensus mechanisms to align with discussed emphasis on integrity and automation in supply chain management.	There is mutual emphasis in both models regarding supply chain management due to the decentralized nature of blockchain.
(Khan, et al., 2018)	This paper presents an Access Control layer that offers a range of authentication methods that include biometrics as well as username/password input. Through validation of these methods, authorized users are allowed to cast their votes, while denying further access to those who fail to authenticate. The focus of this paper lies in ensuring a robust access control mechanism that bolsters privacy and security in the voting blockchain system.	Both models underscore the importance of robust access control and authentication methods. The web-based application feature highlighted in this paper is relevant to *Securi-*Chain's user interface, providing a platform for user interaction and transaction initiation.	*Securi-Chain* has employed a web-based application interface that is discussed in this paper's model, as well as the implementation of transaction initiation.
(Neloy, et al., 2023)	This paper introduces an access control model centred around unique identifiers: voter's State ID, Wallet Address, and Email. This model leverages government databases to validate the identity of voters. Moreover, the access control model ensures that each voter can only cast a single vote. By employing these measures, the paper aims to establish a secure and privacy-enhancing access control mechanism within the voting blockchain system.	Both models mentioned deploy robust access control mechanisms for authentication and access decisions. *Securi-Chain*'s multi-layered model mirrors the reliability and real-time confirmation shown in the paper, allowing for real-time updates to be displayed to the user. Both models can also be adapted for different sectors.	*Securi-Chain* is seen to adopt a multi-layered model that allows for real-time updates to the user interface as well as the ability to be adapted to various domains, such as the model mentioned in the paper.

7. Evaluation

The *Securi-Chain* model, demonstrated through the EMR transfer scenario, exemplifies secure and transparent data exchange in healthcare. Its adaptability and efficiency underline its significance. Central to *Securi-Chain* is its layered architecture, offering a dynamic user interface adaptable across domains for the quick, secure, and ethical transfer of sensitive information. This dynamic adaptability, coupled with an intuitive interface, underscores the model's user-focused approach. The model benefits from the integration of optimal features derived from academic research. This encompasses a robust access control layer (Khan, et al., 2018), a focus on unique identifiers (Neloy, et al., 2023), and an emphasis on granular data privacy (Salonikias, et al., 2022). *Securi-Chain* seamlessly combines these best practices to meet current needs and sets a standard for future initiatives. A deeper examination, comparing *Securi-Chain* to related works, provides valuable perspectives. Khan et al. (2018) highlights the importance of a comprehensive access control layer, advocating for diverse authentication methods, a sentiment echoed in *Securi-Chain*. Neloy et al. (2023) stress the need for unique identifiers to ascertain identity, aligning with *Securi-Chain's* robust access controls. However, the model has its constraints. Its dependence on a layered, real-time framework could lead to delays in larger networks, affecting immediate data access. Its universal suitability might be questionable in areas with infrastructural issues or strict data laws. These challenges indicate areas for future improvement, such as scalability and further adaptability. Salonikias et al. (2022) and Dias et al. (2019) resonate with *Securi-Chain's* focus on data privacy and balancing security and transparency. While Ouaddah et al. (2017) stress transparent access controls in their *FairAccess* framework, *Securi-Chain's* decentralized decision-making enhances its trustworthiness and highlights areas for investigation. Its alignment with key principles from notable research solidifies its credibility. The incorporation of ABAC signifies its versatility. Such mechanisms guarantee data safety and rightful access, mirroring Dias et al. (2019). The layers discussed in *Figure 1* can be seen to prevent potential security issues, such as user misbehaviour. The Audit Layer provides an oversight into user activity that can be viewed by administration. If anything is deemed as misbehaviour, the team will be able to identify where this misbehaviour originated from and from who and will be able to handle this as expected. As blockchain and access control evolve, *Securi-Chain* must adapt to remain scalable. The system's commitment to user-awareness, exemplified by notifying Bob when Alice accesses his EMR, ensures transparency and accountability. With decentralized management, it melds security and responsibility. As *Securi-Chain* evolves, it leads in secure, decentralized systems, but it also acknowledges the need for continuous adaptation in a dynamic cybersecurity environment.

8. Conclusion

In conclusion, the *Securi-Chain* model stands as a paradigm for enhancing the security and privacy of transactions within the realm of blockchain. Its distinct layered architecture, combined with its user-centric design, establishes it as a desirable adaptable solution for access control in blockchain-powered smart contracts. While it has competently integrated features from academic research, the challenges it faces, especially in scalability and adaptability, emphasize the model's potential for future enhancement. By addressing these challenges, *Securi-Chain* not only supports the safe and efficient adoption of blockchain technology across multiple sectors but also emphasizes the crucial balance between confidentiality, integrity, and data accessibility. Moving forward, practical implementations and consistent refinement of the model are essential. Future work in this area should focus on real-world implementation and further refinement of the *Securi-Chain* model, as well as exploring additional use cases and domains for its application in addition to further optimising the scalability and adaptability of *Securi-Chain* in order to test these characteristics for the given architecture of the model.

References

Dias, J. P., Ferreira, H. S. & Martins, A., 2019. A Blockchain-Based Scheme for Access Control in e-Health Scenarios. *Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018),* pp. 238-247.

Ding, S. et al., 2019. A Novel Attribute-Based Access Control Scheme Using Blockchain for IoT. *IEEEAccess,* 7(SPECIAL SECTION ON SECURITY AND PRIVACY FOR CLOUD AND IOT), pp. 38431-38432.

Khan, K. M., Arshad, J. & Khan, M. M., 2018. Secure digital voting system based on blockchain technology. *International Journal of Electronic Government Research (IJEGR),* 14(1), pp. 53-62.

Lashkari, B. & Musilek, P., 2021. A Comprehensive Review of Blockchain Consensus Mechanisms. *IEEEAccess,* Volume 9, pp. 43620-43630.

Liu, H., Deshi, H. & Li, H., 2020. Fabric-iot: A Blockchain-Based Access Control System in IoT. *IEEEAccess,* 8(SPECIAL SECTION ON BLOCKCHAIN-ENABLED TRUSTWORTHY SYSTEMS), pp. 18207-18209.

Liu, J. & Liu, Z., 2019. A Survey on Security Verification of Blockchain Smart Contracts. *IEEEAccess,* Volume 7, pp. 77894-77895.

Maesa, D. D. F., Mori, P. & Ricci, L., 2017. Blockchain Based Access Control. *17th IFIP International Conference on Distributed Applications and Interoperable Systems (DAIS),* pp. 206-220.

Neloy, M. N. et al., 2023. A remote and cost-optimized voting system using blockchain and smart contract. *IET Blockchain.*

Ouaddah, A., Elkalam, A. A. & Ouahman, A. A., 2017. Towards a novel privacy-preserving access control model based on blockchain technology in IoT. *Europe and MENA cooperation advances in information and communication technologies,* pp. 523-533.

Pradnya, P., M, S. & Vidhyacharan, B., 2021. Blockchain for IoT access control, security and privacy: a review. *Wireless Personal Communications,* Volume 117, pp. 1815-1834.

Salonikias, S., Khair, M., Mastoras, T. & Mavridis, I., 2022. Blockchain-based access control in a globalized healthcare provisioning ecosystem. *Electronics,* 11(17), p. 2652.

Yli-Huummo, J. et al., 2016. Where is Current Research on Blockchain Technology? - A Systematic Review. *PLoS ONE,* 11(10), pp. 1-4.

Work in Progress Papers

Using Digital Forensics for Android Smartphone Devices to aid Criminal Investigations

Stephanie Agenbag[1], Andre Henney[1] and Heloise Pieterse[2]
[1]Department of Computer Science, Faculty of Natural Science, University of the Western Cape, Bellville, South Africa
[2]Council for Scientific and Industrial Research, Pretoria, South Africa

3844494@myuwc.ac.za
ahenney@uwc.ac.za
hpieterse@csir.co.za

Abstract: In the past decade, there has been an exponential adoption and ownership of smartphones by billions of users worldwide. However, as smartphone usage increases, criminals have taken advantage of them for illicit or criminal purposes. In criminal investigations, smartphone data has become an invaluable source of information. This study focuses on constructing a snapshot of Android-operated smartphone data to assist investigators in answering critical investigative questions. A thorough review of the literature with regard to the use of smartphone evidence in criminal cases, with the goal of emphasising the investigative phase and the supportive role of mobile data evidence in guiding investigations will be performed. In order to identify additional persons of interest and develop a thorough understanding of the case, the methodology will entail analysing user profiles, smartphone usage patterns, communication logs, application usage, geographic lookups, and device interactions. By conducting a thorough examination of relevant literature, designing a suitable model, and executing a case study, the study intends to offer valuable perspectives on the creation of timelines or visual representations derived from smartphone data. The results of this study will contribute to improving the efficacy of mobile forensics in assisting investigators and facilitating the use of Android-operated smartphone data as supporting evidence in criminal investigations.

Keywords: Digital Forensics, Mobile Forensics, Android-Smartphones, Digital Evidence, Data Analysis

1. Introduction and Background

Smartphones were introduced to the general public over a decade ago. Since their release, they have become increasingly popular over the years as technology evolved. By the end of 2022, there were nearly 6.6 billion smartphone users across the globe (Taylor, 2023). The average smartphone is now a camera, music player, GPS, and telephonic device, all in one. As smartphones have become more popular, they have also become more affordable, making them accessible to a wider range of people.

Smartphones have become an integral part of modern life, but unfortunately, they are also being misused by criminals for various illegal activities. In 2015, Anwar Fisher was convicted of statutory rape and sexual assault based on text messages and inappropriate photos discovered by the victim's mother, resulting in a 2018 confirmation of his six-count conviction (Fisher v S (A51/2016), 2018).

One domain that is largely overlooked is that of criminal investigations, where smartphones have emerged as valuable sources of potential evidence. Acknowledging the strong link between smartphone usage and its relevance in criminal cases, researchers McMillan, Glisson and Bromby (2013) and Zhang et al. (2022) embarked on empirical studies to establish and support this idea. Their findings have inspired the research direction of this study.

In cases where smartphones are seized as evidence, reconstructing their data can be a valuable technique for aiding an investigation. This will give the investigators a method of visualising a timeline of occurrences or a timeline of user activities leading up to an incident. This study will exclusively focus on the Android operating system given Android's dominant global market share of 70.5% in the third quarter of 2023 (Statista Research Department, 2023).

2. Problem Statement

The proposed research will focus on constructing a snapshot of the data retrieved from a smartphone linked to an investigation to assist the investigator. Therefore, the main focus is on the extraction of relevant information from the rich data sources on a smartphone. The goal is to leverage the extracted information to answer critical investigative questions such as who, what, where, when, how, and why certain criminal activities occurred. This information is crucial for building a comprehensive understanding of the activities that

took place on a smartphone. As a result, an overview, either in the form of a timeline or a visual reference, of the user's activities on the smartphone leading up to and during the incident under investigation.

The research question:

How to construct a snapshot of smartphone data to assist a criminal investigation?

3. Literature Review

While the broader field of digital forensics has undergone extensive research, the subdomain of mobile forensics remains in its early stages, marked by a conspicuous scarcity of literature dedicated specifically to mobile forensic investigations. Papers that focused on reviewing mobile forensics in the realm of digital forensics, such as those authored by Dogan and Akbal (2017), and Alatawi et al. (2020), provided a comprehensive overview of the Digital forensic process. This process encompasses four main phases: collection, examination, analysis, and reporting. Additionally, these papers discussed various acquisition methods and tools designed to facilitate the analysis of potential data.

3.1 The Prominence of Mobile Forensics in Criminal Court

McMillan, Glisson and Bromby (2013) took note of the ever-growing popularity of mobile phones and how this could lead to an increase in potential evidence sources in criminal cases. While this correlation appears to be unambiguous, the authors required empirical evidence to substantiate this notion. The study covered cases from 2006 to 2011, obtained through database searches using keywords. Relevant cases were manually selected, and their information extracted into a Microsoft Access database. The analysis of appeal judgments confirms that mobile phone evidence in criminal court cases has been on the rise, showing a correlation between evidence type and the nature of the crimes investigated.

Another Chinese study, drawing inspiration from previous research, investigated various aspects of mobile phone evidence in criminal proceedings from 2013 to 2018. The study focused on determining the proportion of cases that employ mobile phone evidence, examining the types of crimes and mobile phone data involved, exploring the connection between different types of evidence and specific criminal proceedings, and evaluating the significance of mobile phone evidence in court proceedings (Zhang et al., 2022). The study revealed that mobile phone evidence was utilised in a small proportion of the cases included in the dataset, but there was a slight upward trend in the usage of mobile phone evidence over the years.

Despite these papers differing in terms of geographical focus, time period, data sources, and methodologies, they produced very similar results. These papers conclude that it seems like a large amount of mobile phone evidence is filtered out in the stages before court proceedings. Both studies have suggested that future research should focus on exploring the use of mobile data evidence during the investigatory phase rather than solely emphasising its role in court proceedings.

The use of smartphone data as court evidence is a complex and multifaceted subject, with ongoing discussions surrounding the forensic soundness of the techniques and tools employed. However, the aforementioned studies reaffirm the importance of accessing and effectively utilising this data during the investigative process, thereby corroborating the original research direction.

3.2 Extracting and Analysing Smartphone Data

The authors of the paper titled “Mobile Device Forensics: Extracting and Analysing Data from an Android-Based Smartphone” experimented with various methods of extracting and analysing data from Android smartphones using the Sleuth Kit Autopsy (Aziz, Mokhti and Nozri, 2016). The paper begins by highlighting the ubiquity of mobile devices, particularly smartphones, and their increasing use in criminal activities.

The methodology outlines the computer forensic process for Android smartphones, addressing challenges stemming from the diversity of smartphone manufacturers and iterations. The authors utilise the Sleuth Kit Autopsy tool to create a forensic image of the smartphone's data partition. A case study illustrates this approach in a corporate crime investigation, showcasing successful decryption of encrypted messages during the findings and analysis.

In their respective case studies, Jones and Winster (2017) employed Oxygen Forensics, a commercial forensic tool, while Al-Hadadi and AlShidhani (2013) utilised Oxygen Forensics along with UFED Physical Analyzer Cellebrite, mirroring the approach taken in the study conducted by Aziz, Mokhti, and Nozri (2015). Although the depth of analysis in these studies exceeded the capabilities provided by the freeware, the research

conducted by Aziz, Mokhti, and Nozri (2015) shows that retrieving fundamental information and gathering essential data for investigating their case study is achievable. Their paper could have benefited from exploring the case study using both commercial and freeware tools to facilitate result comparisons.

3.3 Exploring Potential Sources of Evidence in Smartphone Data

The research paper titled "Mobile Forensics: Beyond Traditional Sources of Digital Evidence" explores the importance of contemporary sources of data on mobile devices for digital forensic investigations (Pieterse, 2020). Traditionally, digital forensic investigations have focused on data such as contacts, messages, call history, and web browsing. However, this paper highlights the often-overlooked sources of data, including log files, usage statistics, and event data. The paper presents the Pre-Analysed Device Snapshot (PADS) model for creating a snapshot of a mobile device's state at the time of acquisition using contemporary data sources. It was practically evaluated with data from a Samsung Galaxy S5 Mini running Android 6.0.1, successfully identifying the user profile, recent app usage, and other activities on the device.

In principle, expanding the scope of what we can investigate offers investigators access to a multitude of potential data. However, the practicality of this approach is hindered by the substantial increase in the storage capacity of smartphones and the typically high number of installed apps. The enormous time, financial, and resource commitment that such a thorough investigation could require is the cause of this formidable challenge. This is not addressed in the paper would necessitate investigation if the model or any derivative form of inspiration were to be incorporated into the proposed study.

4. Research Framework

A qualitative research approach will be used to delve into the complex interactions and behaviours within smartphone data during criminal investigations. This method provides in-depth insights into the experiences, perspectives, and contextual details of individuals involved.

The study will employ an in-depth literature review, which will specifically focus on the following: (1) available smartphone data and (2) acquisition methods to extract the data from the smartphones. The objective is to develop a thorough understanding of the most advanced techniques and data sources that are currently available for smartphone data.

Thereafter, based on the extracted features, a model will be developed to construct a snapshot of the smartphone data. The model will accept as input the smartphone data and as output produce the timeline or visual reference that presents the user's activities leading up to and during the incident.

Finally, the model's evaluation will involve a case study focusing on specific criminal investigations, which may use either simulated or real-world smartphone data. If acquiring real-world smartphone data is not feasible, the authors will simulate data sources based on various smartphone usage scenarios. Subject to ethical and legal considerations, this can be obtained from sources like law enforcement authorities. Ethical considerations and data protection will be crucial if the case study uses real-world data.

The results will be addressed in relation to the research question and aims, with a focus on how investigators using smartphone data in criminal cases can use them in real-world situations. The research framework outlined above will form the groundwork of the proposed study that will provide valuable insights into generating timelines from smartphone data, contributing to the existing knowledge in this domain.

5. Contribution and Conclusion

The motivation for the proposed study stems from the conclusions drawn in the studies by McMillan, Glisson and Bromby (2013) and Zhang et al. (2022). The findings of both studies indicate that rather than solely emphasising its role in court proceedings, future studies should investigate the application of mobile data evidence during the investigative process. Thus, the research aims to answer critical investigative questions that will guide the investigator such as who, what, where, when, how, and why certain criminal activities occurred.

- The study aims to identify other persons of interest for the investigator to determine who communicated with the device.
- The research will focus on analysing the most relevant smartphone data for the investigator.
- Determining the smartphone's location at different times is challenging due to privacy concerns, therefore alternative techniques will be used to mitigate privacy concerns.

- The research aims to identify periods of activity or inactivity on the smartphone by investigating detailed application activity.
- The study will identify other devices of interest, including external connections via Wi-Fi and Bluetooth.

The model that will be developed will draw inspiration from the PADS model, introduced by Pieterse (2020), and aims to assist the investigator in quickly and more effectively constructing a timeline or rather snapshot by using both traditional and contemporary data sources to reconstruct smartphone data.

Reconstructed smartphone data can include but is not limited to:

- location data, which can assist investigators in identifying relevant evidence sources, such as surveillance footage, eyewitness accounts, and other devices in proximity at the time of the incident.
- digital footprints, such as online search history and social media activity. This can guide investigators in identifying relevant evidence sources, such as websites visited, or social media accounts used to communicate with individuals involved in the incident.

The results of this study will contribute to the growing literature and research in this field. Additionally, it will enhance the efficacy of mobile forensics in assisting investigators and facilitating the use of Android-operated smartphone data as supporting evidence in criminal investigations.

References

Alatawi, H., Alenazi, K., Alshehri, S., Alshamakhi, S., Mustafa, M., & Aljaedi, A. (2020). “Mobile Forensics: A Review”, in 2020 International Conference on Computing and Information Technology, ICCIT 2020.

Al-Hadadi, M., & AlShidhani, A. (2013). “Smartphone Forensics Analysis: A Case Study”, International Journal of Computer and Electrical Engineering, pp. 576–580.

Aziz, N.A., Mokhti, F. and Nozri, M.N.M. (2016) “Mobile Device Forensics: Extracting and Analysing Data from an Android-Based Smartphone”, in Proceedings - 4th International Conference on Cyber Security, Cyber Warfare, and Digital Forensics, CyberSec 2015. Institute of Electrical and Electronics Engineers Inc., pp. 123–128.

Dogan, S., & Akbal, E. (2017). “Analysis of Mobile Phones in Digital Forensics”, in 40th International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2017, pp. 1241—1244.

Fisher v S (A51/2016) (2018), [online], https://www.saflii.org/za/cases/ZAWCHC/2018/15.html

Jones, G., Winster, S. (2017). “Forensics Analysis On Smart Phones Using Mobile Forensics Tools”, International Journal of Computational Intelligence Research, Vol 13, No.8, pp 1859–1869.

McMillan, J.E.R., Glisson, W.B. and Bromby, M. (2013) “Investigating the increase in mobile phone evidence in criminal activities”, in Proceedings of the Annual Hawaii International Conference on System Sciences, pp. 4900–4909.

Pieterse, H. (2020) Mobile forensics: “Beyond traditional sources of digital evidence”, in European Conference on Information Warfare and Security, ECCWS. Curran Associates Inc., pp. 295–303.

Statista Research Department (2023) “Market share of mobile operating systems worldwide 2009-2023”, Statistica. [online], https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/

Taylor, P. (2023) “Mobile network subscriptions worldwide 2028”, Statistica. [online], https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

Zhang, A. et al. (2022) “Investigating the uses of mobile phone evidence in China criminal proceedings”, *Science and Justice,* Vol 62, No. 3, pp. 385–398.

Deep Learning-Based Framework for Detecting Malicious Insider-Inspired Cyberattacks Activities in Organisations

Gibson Chengetanai[1], Tendai Chandigere[1], Pepukai Chengetanai[2] and Rachna Verma[1]
[1]School of Computing and Information Systems, BAC, Gaborone, Botswana
[2]School of Accounting and Finance, GIPS, Gaborone, Botswana

gibsonc@bac.ac.bw

Abstract: Cyberattacks are happening at an alarming rate both in developed and developing countries. This is due to more users now being connected to the global village (internet). Significant strides have been taken by organisations to protect information technology assets together with data, by doing defense-in-depth, using firewalls and access control approaches collectively. These approaches work well in detecting attacks by outsider cyber-attackers. In recent cyberattacks the perpetrators have been those within the organisation, as they can easily bypass security measures especially those with high privileges and they can go undetected for quite a long time. We propose a deep learning approach termed Automatic_ IDS_ Deep model (framework) that is infused with intrusion detection systems to give timely detection of malicious activities by those within the organisation. Experiments were conducted and averaging of results was done to determine accuracy, recall, and precision of the proposed model. The model (framework) offers better results on its performance in detecting attacks that are perpetrated within the organisation.

Keywords: Deep learning, Intrusion detection system, Cyberattacks, Malicious users

1. Introduction

Cyberattacks are happening every day on a larger scale as more and more people get connected to the internet. Users in the workplace or in their homes, are all now going online increasingly on social networking sites such as X formerly known as Twitter, Facebook, TikTok among the many such web services available these days. Employees in organisations can divert some of their time to network with fellow professionals, friends, and other people around the globe (Smitha, Siano and Parenta, 2023; Lo et al, 2018).

Organisations through their IT specialists are putting measures to detect malicious attacks attempts towards the IT infrastructure sand its associated data and systems in a timely manner. Despite corporate attempts to protect valuable IT assets, cyber-attacks are still occurring, though business entities do not divulge these attempts for fear of reputational damage. It has been observed that many of the attempts to prevent attacks are effective for people outside the organisation. Insiders within organisations (usually disgruntled employees), and contract vendors can pose major a threat in the organisation (Insider threats[6]; Smitha, Siano and Parenta, 2023). Cyberattacks are due to the following (Figure 1):

Financial gain - employees can steal corporate data and sell it to the organization's rivals (Fujii, Kurima and Isobe, 2019). Hackers, for example can disable accounts and create new accounts once they gain privileged access to systems. Hackers can copy data and use that copied data which may include pricing models that have been used in the organisation and some can even go with the corporate data when they leave the organisation (Smitha, Siano and Parenta, 2023; Lo et al, 2018).

Theft of intellectual property – this usually occurs in state sponsored entities, whereby governments spy on each other on areas of research and development. Around 44% of cyberattacks are around intellectual property theft (Inayat et al., 2022).

Disgruntled employees and organizational politics can all contribute to individuals stealing corporate data as they are unhappy with the status quo in the organisation. Figure 1 shows the main motivation for attacks perpetrated within the organisation (Smitha, Siano and Parenta, 2023).

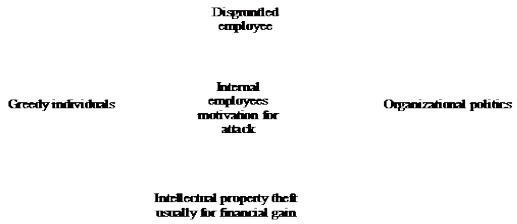


Figure 1: Motivations for insider cyberattacks in organisations

Organizational politics can lead to some employees wanting to punish the employer as they are disgruntled by how things are running in their workplace (Modini et al, 2020). This could be because such people have access rights to the system and can easily evade detection. Greedy individuals can be individuals who want to make money by selling data to competitors for a fee (Saxena et al., 2020).

2. Related Work

Machine learning approaches have been used to help in detecting malicious events in IT systems in organisations. Machine learning approaches help in detecting some abnormal trends with high accuracy. Most of these machine learning approaches are deployed on host or network perimeters to help in detecting an anomaly based on the signature-based approach. Machine learning approaches that have been used to detect anomalies in IT systems include support vector machines (SVM). SVM works in multiclass environment, and they are good at solving linear problems (Saxena et al., 2020; Modini et al., 2020). SVM approach does not scale well on huge data sets whilst artificial neural networks (ANNs) are prone to overfitting (Weng, Li and Zhu, 2020), and it requires more time for model training as compared to other machine learning approaches (Liu, 2021; Lang, 2019) . Naïve bayes is a supervised machine learning based approach that work based on Bayesian theorem and its advantage is of simplicity and easy of computation (Al-Shehariand Alsowail, 2021; Al-Mhiuai et al, 2020). K means clustering works by discovering clusters in the data object with k being number of clusters that have been formed (Valliammal and Shaju, 2018; Gayathri, Sajanhar and Xiang, 2020). When clustering is done, objects in a particular cluster will have similar characteristics. Logistics regression (LR) LR works well on linear data and thus fails to work well on non-linear data sets (Smitha, Siano and Parente, 2023). Decision trees are another machine learning method, however its weaknesses is that it does not consider correlation in decision making (Lo et al, 2018; Gayathri, Sajanhar and Xiang, 2020).

Anomaly based detection methods check on user behaviour and where there is a deviation from the expected behaviour, alert messages are sent to the administrators (Weng, Li and Zhu, 2020). The machine learning approaches each have some weakness in using them since most of them are associated with high false positive alert messages (Gayathri, Sajanhar and Xiang, 2020). From literature it is evident that machine learning approaches to finding malicious activities perpetrated within the organisation have a weakness of giving less accuracy. This can be attributed to machine learning approaches' inherent use of small data for training and testing data sets. A deep learning method is proposed which is a subset of machine learning. Deep learning approach was chosen mainly because it provides better accuracy relative to machine learning approaches since it uses huge data during model training.

3. Proposed Automatic_IDS_ Deep Learning Framework

Our contribution is as follows:

- Development of a deep learning approach called Automatic_IDS_Deep framework which takes into consideration a number of activities that are extracted from log files such as logon and logout times, devices used to connect, http sites visited, event log relating to removable devices such as flash memory and USB devices connected to the systems and times they have been connected, communication patterns, printer activity logs, database audit logs and access logs.
- New algorithm coupled with experimental results and evaluations to determine the accuracy of the proposed Automatic_IDS_Deep model.

Our proposed framework is based on user behaviour compilation looking into activities that users do on the system on a daily basis. These include checking and extracting information such as logon times, daily system logs including systems logs, web server logs and database audit logs. Event logs together with the device used to logon to the system and traffic send with particular attention to actions such as failed events from various IP addresses will be recorded. When activities are recorded, preprocessing is done to remove any redundancies, i.e. during the data cleaning process. Missing values are replaced with average values and an algorithm is infused into the system that does preprocessing. After data cleaning, feature extraction is done to determine on what is normal considered normal or otherwise, based on historical data. On feature extraction the following Table 1 shows the feature set variables that will be recorded to determine malicious users.

Table 1: Features to be selected when profiling inside user (internal employees) activities to determine malicious ones

Domain	Feature	Feature set to be observed
Network	HTTP	URLs, browser type, downloads
	Email	Source address, destination address, time, attachments
	TCP/IP	Source and destination IP addresses, size of data packets, data traffic size
Host	File and folders	Check on open, create, copy, move, delete operations
	USB	Device ID, duration, and associated actions on the device such as copy, delete, or paste
	System call	Process name and ID, registry, and file system
	Login or logout operations	User ID, timestamp, login duration, login frequency and logout time

After feature selection is done, we train and test data on the proposed model and evaluate its performance. Our proposed solution offers better solution of reducing false positives on alert messages because it considers all activities that users can do. The algorithm below shows the Automatic_IDS_Deep Learning based framework for detecting malicious events in the workplace.

Algorithm 1. Deep learning-based algorithm for detecting malicious events in the workplace

```
INPUT:    Check on data coming from different sources on the
            network β= {logon log, Database log, file log, print
             log, web server log, error log, file access log, TCP/IP,
              flow log, events log}
               iteration (t)
OUTPUT:   Network status on any cyberattacks showing machines
            that have been affected by malicious events
           Classification for malicous and non-malicious actions for review by network administrators
Procedure
Step 1   iteration:=0
Step 2: examine data packets in the network
      // data logs can be coming from logon, database file, print log,
      web server, error log, file access, TCP/IP flow log, events log
Step 3: pre-process using the configured algorithm
Step 4:  use feature selection algorithm
Step 5:  Training the model    //extract relevant features from data sets and train model with Epochs=200
          if result point to malicious activities then
          add the workstation or mobile nodes that has such
               activities
         else   //No malicious activities on the network
             save the workstation or mobile terminal as correct node
         end if
Step 6:  Evaluate the performance of the model
Step 7: Display the results on malicious nodes as generated in step 5
      t=t+1
end procedure
```

Organisations can also have the log file duplicated to trusted offsite parties that are credible. This will help in detecting malicious activities that might be done by administrators who might have privileges. These privileges include deleting data/evidence that might be needed to detect the culprits inside the organisation as they have capabilities for deleting user activities in the log file(s).

4. Experimental Results and Evaluation

In this section we present results showing the performance of the proposed Automatic_IDS_Deep model. The data set that was used to determine insider threat was synthetic data obtained from CERT, a project at Carnegie University. The items that make up the data set include login and logout, email and web, USB, system calls. The data set was generated using two processes to simulate an attack where malicious software can be injected into the system using removable devices such as memory sticks and lastly simulation using email attachments was done and injecting data to the Windows server 2019. Performance metrics that were explored were accuracy, recall, precision, and F1-score. The computation of each performance metric is below.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (1)$$

The F-measure of harmonic mean is computed as:

$$F1 - score = \frac{2 * precision * recall}{precision + recall} \quad (2)$$

with precision being computed as

$$Precision = TP/(TP + FP) \quad (3)$$

and recall also known as sensitivity computed as

$$Recall = TP/(TP + FN) \quad (4)$$

with TP being true positive that is an activity identified by the system as actually being malicious, TN is true negative being an activity that has been correctly identified as non-malicious and is actually non-malicious and lastly FN is for false negative which is an activity that is not detected by the system as malicious but is actually malicious

The results on Table 2 shows the performance of the proposed Automatic_IDS_Deep Model.

Table 2: Experimental results for the Automatic_IDS_Deep Learning framework

Training Accuracy	Training loss	Validation Accuracy	Precision	Recall	F1-Score
0.96	0.18	0.97	0.93	0.9	0.914

5. Conclusion

Cyberattacks in organisations have been observed to come from employees within the organisation. Internal employees who have privileged access to information that resides in the organisation can copy corporate data/information and give this information to the organization's competitors for a fee. This research has managed to come up with a framework , termed Automatic_ IDS_ Deep Framework. The model is novel in the sense that it infuses intrusion detection and combines it with deep learning which is based on convolutional neural networks. Synthetic data from Carnegie Mellon University was used to test our proposed model based on deep learning. Since most of the malicious activities happening in organisations are done by inside employees as seen from literature, we proposed a secure deep learning model called Automatic_ IDS_ Deep model. The adoption of this framework can help in mitigating attacks timeously by sending alert messages to administrators in the organisation for action. In future, we intend to do some comparison of our proposed model with other machine learning models such as (Naïve Bayes, SVM, etc.) and other deep learning models to check on the performance of our model relative to others.

References

Al-Shehari, T. and Alsowail, R.A., 2021. An Insider Data Leakage Detection Using One-Hot Encoding, Synthetic Minority Oversampling and Machine Learning Techniques. Journal of Entropy, Vol. 23 No. 10, pp.1-24

Fujii, S., Kurima, I. and Isobe, Y., 2019. Scoring Method for Detecting Potential Insider Threat based on Suspicious User Behavior using Endpoint Logs. Athens: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), pp.1-7

Gayathri, R.G., Sajjanhar, A. and Xiang, Y., 2020. Image-Based Feature Representation for Insider Threat Classification. Applied Sciences, Vol .10 No. 14, pp4945-4953

Gheyas, I.A. and Abdallah, A.E., 2016. Detection and prediction of insider threats to cyber security: a systematic literature review and meta-analysis. Big Data Analytics, Vol .1 ,pp.1-29

Hammoudeh, M., Watters, P., Epiphaniou, G., Kayes, A.S.M. and Pinto, P., 2021. Special Issue "Security Threats and Countermeasures in Cyber-Physical Systems". Journal of Sensor and Actuator Networks, Vol. 10 No. 3, pp.1-4

https://techjury.net/blog/insider-threat-statistics/#gref Accessed 7 January 2024

Inayat, U., Muhammad, F.Z., Mahmood, S., Khalid, H.M. and Benbouzid, M., 2022. Learning-Based Methods for Cyber Attacks Detection in IoT Systems: A Survey on Methods, Analysis, and Future Prospects. Electronics, Vol. 11 No. 9, pp.1-42

Lang, B., 2019. Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey. Applied Sciences, Vol. 9 No. 20, pp.43-76

Liu, H., 2021. An insider threat detection system based on user and entity behavior analysis. Journal of Physics: Conference Series, Vol. 1999 No. 1, pp.1-11

Lo, O., Buchanan, W.J., Griffiths, P. and Macfarlane, R., 2018. Distance Measurement Methods for Improved Insider Threat Detection. Security and Communication Networks, pp.1-19

Modini, J., Vanzomeren, M., Fowler, S., Joiner, K. and Lynar, T., 2020. Rising to the Challenge of Insider Threats for Middle Powers. Reading: Academic Conferences International Limited, pp.1-6

Mohammed Nasser Al-Mhiqani, Ahmad, R., Z., Yassin, W., Hassan, A., Karrar, H.A., Ali, N.S. and Yunos, Z., 2020. A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations. Applied Sciences, Vol. 10 No. 15, pp.1-41

Saxena, N., Hayes, E., Bertino, E., Ojo, P., Kim-Kwang, R. and Burnap, P., 2020. Impact and Key Challenges of Insider Threats on Organizations and Critical Businesses. Electronics, Vol. 9 No 9., pp.1-29

Smitha, J.P., Siano, P. and Parente, M., 2023. Review of Cybersecurity Analysis in Smart Distribution Systems and Future Directions for Using Unsupervised Learning Methods for Cyber Detection. Energies, Vol. 16 No. 4, pp.1-24

Valliammal, N. and Shaju, B., 2018. Deep learning algorithm based cyber-attack detection in cyber-physical systems-a survey. International Journal of Advanced Technology and Engineering Exploration, Vol. 5 No. 49, pp. 489-494.

Weng, W., Li, W and Zhu, L. (2020). "Enhancing Medical Smartphone Networks via Blockchain-Based Trust Management Against Insider Attacks". IEEE Transactions on Engineering Management, Vol. 67, No. 4, pp.1377-1386

Active Gait System for Real-Time Surveillance Against Cyber-Physical Attacks

Glen Lehlohonolo Moepi and Topside Ehleketani Mathonsi
Department of Information Technology, Tshwane University of Technology, Pretoria, South Africa

MoepiGL@tut.ac.za
MathonsiTE@tut.ac.za

Abstract: Cyberterrorism, espionage, and warfare are serious threats to national security. These attacks can harm people or destroy critical infrastructures like the data centres, computer networks, and systems. Surveillance systems currently used in monitoring critical infrastructures, national key points, and military exclusion zones (MEZ) are ineffective in detecting unauthorised intrusions. These issues compromise the stability of the countries, and the safety of the citizens and result in the loss of important assets. This experimental research study developed a Cyber Physical Security (CPS) defense gait-recognition monitoring system. Autonomous Machine Learning (ML) technology was employed to enhance the precision and reliability of the system against CPA, in tracking access, managing security clearances, and triggering alerts in the event of unauthorized entries to restricted areas.

Keywords: Cyber Physical Security (CPS), Deep Learning (DL), Gait analysis, Gait event detection, Machine Learning (ML)

1. Introduction

Power grids, necluear plants, and critical computer networks are currently facing the challenges of physical attacks. A study by Wlazlo *et al.* in (2021) and (Tyagi & Sreenath, 2021), shows that Cyber-Physical Attacks (CPA) are on the rise. Token-based and knowledge-based identification schemes are commonly used for verification when granting entry and security clearance to restricted areas. However, physical tokens can be cloned and are also susceptible to theft. In mitigation to that, the introduction of biometric systems can be used for security in authorization and intrusion detection systems (Robb, 2022).

Moro *et al.* (2022), compared Markerless and Marker-Based Gait Analysis, and it was determined that gait-recognition technology exhibits a high level of reliability and efficacy in accurately identifying individuals, achieving an accuracy rate of up to 99%. A different study by (Lin *et al.,* 2023) proved that the accuracy and dependability of systems used for authorizations can be increased by utilising gait recognition technology. The gait recognition technology in conjunction with Artificial Intelligence (AI) has been used successfully in surveillance (Yan *et al.*, 2021).

To reduce the risk of CPA and efficient administration of authorizations, this paper proposes an active gait recognition system using (ML). The paper seeks to provide a highly accurate and secure method of tracking authorized and detecting un-authorized individuals by using unique gait patterns. This paper endeavours to make a sizeable contribution to the domains of access and authentication systems.

The following is the order in which this paper is organized: Section 1, introduces gaits, AI and ML applications, the objectives of the study, the problem being addressed, and the paper's contributions. Section 2 highlights the related work on gaits systems. In Section 3, this paper experimented with the use of the proposed gait system. Section 4, brings the paper to a close with the results analysis and makes recommendations for future research.

2. Related Work

A research study by (Álvarez-Aparicio *et al.,* 2022) investigated the use of gait recognition technology as a means of biometric recognition. The study found that gait recognition technology can achieve high levels of accuracy, with an equal error rate of 1.4%. The study also showed that gait recognition technology can be used in real-time scenarios, making it a reliable option for authorization and intrusion detection. Similar to the study, this paper used the active gait recognition system and for real-time authorization and intrusion detection system (IDS) that will trigger alarms upon the detection of anomalies, a probable deterrent to intruders.

A surveillance detection study by (Esan et al., 2023), proved that Deep Learning (DL) technique model has 95% accuracy in anomaly detection. Their study combined Convolutional Neural Network (CNN) and Long Short-Term Memory (LSM). Another study by (Albuquerque *et al.,* 2021), also proved that the use of AI and gaits technology is effective in wide areas and long distances object identification. Similar to this study, DL is a fundamental element in object learning and anticipating normal gaits actions in high-risk facilities and effective management

of assets. Overall, the gait recognition technology has the potential to revolutionize authorization and intrusion detection systems.

Another study by (Lee & Tang, 2021), proposed a portable system for gait phase detection using AI. The system was designed to be used in healthcare settings, but the technology can be applied in access systems. The study showed that the proposed system achieved an accuracy of 91.3% in gait phase detection. Another embedded system was designed by (Pradel *et al.,* 2019) to analyze gait patterns and provide feedback to patients and healthcare providers. These studies showed that the systems were able to accurately measure gait parameters. The embedded gait analysis system proposed by (Pradel *et al.,* 2019) could be used to identify abnormal gait patterns in employees and livestock, which could be an indication of injury or fatigue; or animal is under attack. However, using embedded systems, digital wearables, and accelerometers is costly for a with many assets to track and monitor. This paper seeks to develop a system that is not cost-prohibitive, using a standard video camera and unsupervised clustering DL techniques.

3. Proposed Solution

The identity verification is implemented as part of the gait recognition using five different authentication sets. Each authorization is based on a set of following processes according to the technique. The proposed protocol includes several elements, such as:

- Acquiring gait characteristics.
- Extraction and processing of gait features.
- Reduced feature count.
- ML methods for classification.
- Depending on the outcomes of the categorization.

These processes enable reliable authentication that improves the gait-recognition system's security and usability as illustrated in Figure 1.

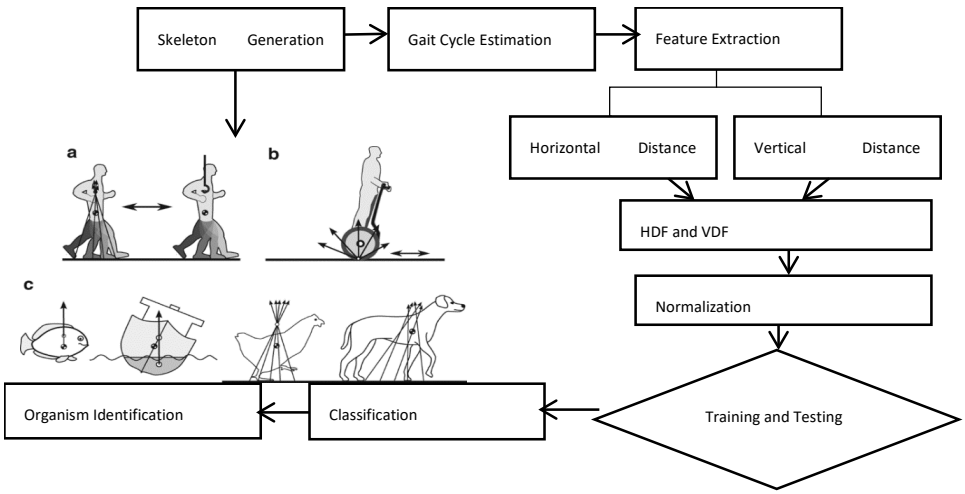


Figure 1: Active Gait System Architecture Design Incorporated (Maus, *et al.,* 2012)

4. The Implementation of the Proposed Scheme

A controlled environment setting for experimentation was established. A format for trial-and-error testing was employed with the Intersection over Union (IoU) ratio, as a cutoff point for assessing the object-detection models' accuracy. The region of overlap between the ground reference bounding box and the anticipated bounding box is the numerator. The area of union, or the region covered by both bounding boxes, served as the denominator.

Stage1: Webcam Object Detection

YOLOv4 is extremely effective in object identification technology. As depicted in Figure 2, YOLOv4 uses a deep Convolutional Neural Network (CSPDarknet53), a more sophisticated backbone network. With the help of this network design, YOLOv4 is better able to properly recognize objects by extracting more specific and discriminative information from the input image. Additionally, YOLOv4 uses several optimization approaches to boost its accuracy and speed. A good example of this is the incorporation of cross-stage partial connections

(CSP), which increase data flow and make it easier to interchange both high-level and low-level characteristics, improving detection performance. In order to properly handle objects of various sizes, YOLOv4 additionally makes use of the PANet (Path Aggregation Network) and Spatial Pyramid Pooling (SPP) modules.

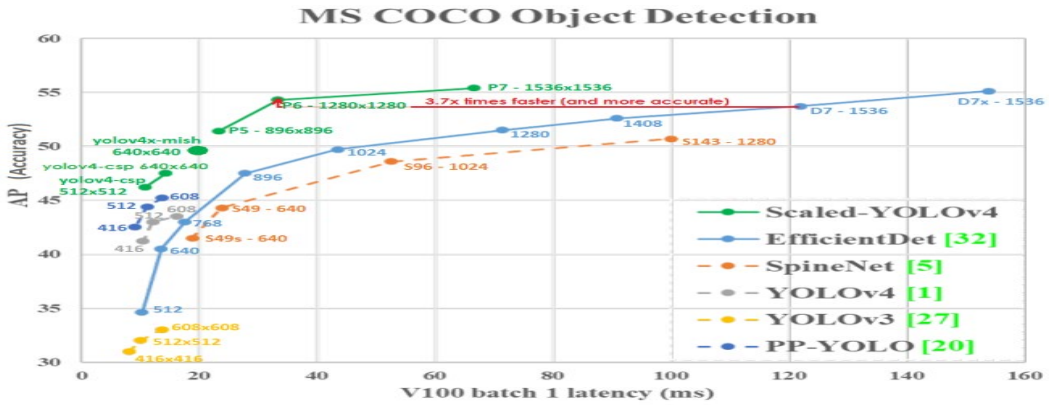


Figure 2: Object detection

1. Import Dependencies

Importing dependencies allows access and use of external libraries, modules, or frameworks that provide the necessary functionality and tools. The importation of specific dependencies serves two main purposes: enabling the execution of YOLOv4 object detection and leveraging the capabilities of Google Colab. To perform YOLOv4 object detections on a webcam, necessary learning dependencies include deep learning frameworks such as TensorFlow or PyTorch, which provide the necessary tools for building and training neural networks. Google Colab is a cloud-based platform that offers a hosted Jupyter notebook environment with access to computational resources, including GPU acceleration.

2. Cloning and Setting up a Darknet for YOLOv4

Cloning and setting up Darknet for YOLOv4 using AlexeyAB's darknet repository provides the necessary framework for implementing and running YOLOv4 object detections. AlexeyAB's darknet repository is ideal for its comprehensive implementation of YOLOv4 and its compatibility with different hardware platforms.

3. Adoption of Darknet for Python

Darknet for Python involves importing and utilizing pre-built functions from the darknet.py module. This allows seamless integratation of YOLOv4 functionality into the Python code, to perform object detections and work with the YOLOv4 model efficiently.

4. Helper Functions

Helper functions for image type conversion ensures compatibility and consistency between different image formats.

4.1 The Gait Recognition System Function Outputs

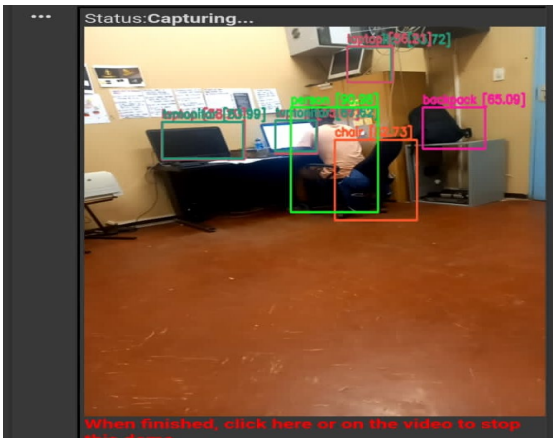


Figure 3: Object Detection YOLOv4

Using Yolo each frame of a video is treated like a picture and objects are detected at each frame.

5. Results and Analysis

The integration of YOLOv4 and Darknet for Python improved the accuracy of the gait-recognition system as seen in Figure 3. This paper observed a 20% enhancement in accuracy for monitoring employees and livestock, while intruder detection accuracy improved by 15%. The integration of Roboflow AI training software reduced training time by 30%, contributing to a faster and more efficient AI model. In the 'findGirl' project as seen in the below Figure 4, the system achieved a remarkable 90% accuracy rate in detecting the 'Girl' within a crowded environment. Roboflow utilized YOLOv8 in later stages, which contributed to the exceptional results achieved in the 'findGirl' project. The system achieved a 90% accuracy rate and maintained a high 85% accuracy rate both indoors and outdoors.

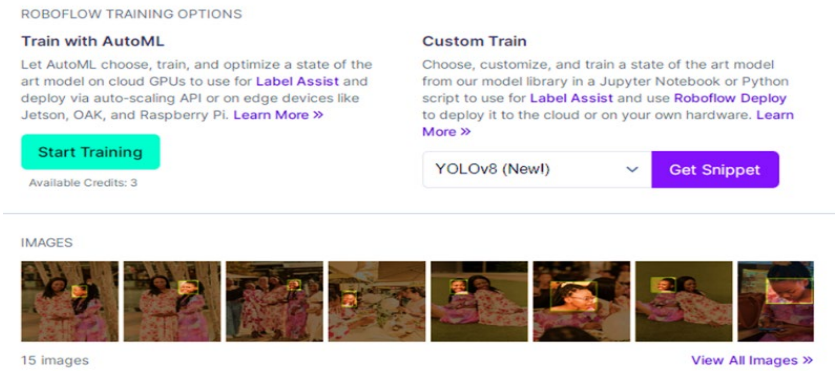


Figure 4: Roboflow AI Detection using YOLOv8

Table 1: Tabulated Detection Results

Aspect	Results
YOLOv4 Integration with Darknet	Improvement of 20% in gait-recognition accuracy. Improvement of 15% in intruder detection accuracy.
Roboflow AI Training Software (YOLOv8)	Reduction of 30% in training time.
'findGirl' Project (YOLOv8)	Accuracy of 90% in detecting 'Girl' within a crowded environment.
Webcam Object Detection (YOLOv4)	Accuracy of 85% in object detection both indoors and outdoors.

6. Conclusion and Future Work

This paper developed an authorization and intrusion detection system using active gaits recognition and ML technology to reduce CPA on critical infrastructures. The active gait system was deployed in an experimental field testing environment. The results demonstrated an acceptable level of usability, improved security, and the ability to learn and adapt to different environments. The ultimate goal of the research study was partly accomplished, and since this is still a work in progress, more investigations will be undertaken to explore related tools and updated techniques related to the study.

References

Albuquerque, P., Verlekar T. T., Correia P. L. & Soares L. D., 2021. A spatiotemporal deep learning approach for automatic pathological gait classification: Sensors, doi:10.3390/s21186202,21(18):6202.

Álvarez-Aparicio, C., Guerrero-Higueras, Á. M., González-Santamarta, M. Á., Campazas-Vega, A., Matellán, V., &

Esan, O. A., Esan, D. O., Mbodila, M., Abiodun, F., & Koranteng, K., 2023. Surveillance detection of anomalous activities with optimized deep learning technique in crowded scenes. Bulletin of Electrical Engineering and Informatics,doi:10.11591/eei.v12i3.4471,12(3):1674–1683.

Fernández-Llamas, C., 2022. Biometric recognition through gait analysis. Scientific Reports, doi:10.1038/s41598-022-18806-4, 12(1):1-10.

Lee, K., & Tang, W. ,2021. A fully wireless wearable motion tracking system with 3D human model for gait analysis. Sensors, 4051. doi:10.3390/s21124051, 21(12):1-11.

Lin, Q., Xu, X., Zhang, F., Xian, D., Yao, K., Zhao, N., … JIANG, Z., 2023. Human wearable gait recognition system based on flexible and distributed pressure sensor. Measurement, 113726. doi:10.1016/j.measurement.2023.113726, 1-25.

Maus, H.M., Lipfert, S., Gross, M. et al. Upright human gait did not provide a major mechanical challenge for our ancestors. Nat Commun 1, 70 (2010). https://doi.org/10.1038/ncomms1073.

moro, m., marchesi, g., hesse, f., odone, f., & casadio, M. ,2022. Markerless vs. marker-based gait analysis: A proof of concept study. Sensorsdoi:10.3390/s22052011, 22(5).

Pradel, G., Li, T., Pradon, D., & Roche, N., 2019. An embedded gait analysis system for CNS injury patients. Assistive and Rehabilitation Engineering. doi:10.5772/intechopen.83826, (1):2-24.

Robb, D., 2022. The future of biometrics in the workplace, SHRM. Available at: https://www.shrm.org/resourcesandtools/hr-topics/technology/pages/the-future-biometrics-workplace.aspx.

Tyagi, A. K., & Sreenath, N., 2021. Cyber Physical Systems: Analyses, challenges and possible solutions, Internet of Things and Cyber-Physical Systems, 21(1):22-33, ISSN 2667-3452.

Wlazlo, P., Sahu, A., Mao, Z., Huang, H., Goulart, A., Davis, K., & Zonouz, S., 2021. Man-in-the-middle attacks and defence in a power system cyber-physical testbed. IET Cyber-Phys. Syst., Theory Appl. 6(3):164-177.

Yan, S.H., Liu, Y.C., Li, W., & Zhang, K., 2021. Gait phase detection by using a portable system and Artificial Neural Network. Medicine in Novel Technology and Devices, 12, 100092. doi:10.1016/j.medntd.2021.100092,1-8.

Utilization and Sharing of Cyber Threat Intelligence Produced by Open-Source Intelligence

Jyri Rajamäki and Stephen McMenamin
Laurea University of Applied Sciences, Espoo, Finland

jyri.rajamaki@laurea.fi
Stephen.McMenamin@student.laurea.fi

Abstract: Open-source intelligence (OSINT) is crucial for enhancing organizational cybersecurity by proactively identifying and mitigating potential threats using publicly available information. This study, part of the DYNAMO project, explores the production of cyber threat information (CTI) through OSINT, its application in safeguarding against cyber threats, and the necessary elements for secure information exchange between organizations. The authors employed an integrative literature review of various sources, including industry literature, articles, blog posts, studies, and organizational websites, which were then systematically analyzed using content analysis. The research focuses on OSINT tools and techniques emphasizing the need for expertise in discerning relevant data and respecting privacy rights. Human judgment is highlighted as crucial in ethical decision-making despite the significant role of technology in data collection. Platforms like the Malware Information Sharing Platform (MISP) facilitate the sharing of threat information, promoting prevention and identification of cyber-attacks. Ethical considerations, adherence to data protection legislation, and compliance with directives like the revision of the Network and Information Security Directive (NIS2) and artificial intelligence regulations are paramount. In conclusion, OSINT is a valuable tool for cybersecurity, requiring expertise, transparent processes, and a balanced integration of technology and human skills. The ethical dimensions of OSINT and the role of artificial intelligence merit separate in-depth studies.

Keywords: OSINT, Cyber threat intelligence, Cyber information sharing, DYNAMO project

1. Introduction

Open-source intelligence (OSINT) is conducted by various entities, including law enforcement agencies, defense forces, investigative journalists, and cybersecurity professionals. Malicious actors, such as cybercriminals or state-sponsored threat actors, also leverage OSINT techniques. The objective is to gather information about organizations or individuals that can be exploited in cyberattacks or pose a threat to security. This information is utilized in cyberattacks, prevention efforts, criminal investigations, or to support business strategies (SANS Institute 2023; Fruhlinger, Sharma & Breeden 2023).

OSINT is, in many ways, a counterpart to Operational Security (OPSEC), a process through which organizations protect information that should not be made public. For example, an employee's social media post may inadvertently contain sensitive information about the organization, or IoT devices connected to the public network may reveal information beneficial to threat actors regarding the organization's technologies. OSINT enables the proactive identification of potential threats and suggests measures to mitigate them (Fruhlinger et al. 2023; Rose 2020).

The DYNAMO project, funded by the European Union, aims to create a platform to reduce cyber threats to critical sectors such as energy, health, and transportation infrastructure. The platform supports all stages of cyber resilience - preparation, prevention, protection, response, and recovery - tailored to the needs of critical infrastructure. Its goal is to enhance the security of organizations and society (DYNAMO 2023; Packham 2022).

One component of the platform involves generating Cyber Threat Intelligence (CTI) through Indicators of Compromise (IoC), enhancing an organization's information and cyber monitoring capabilities (DYNAMO 2023). IoC are clues or observations of threatening activities, such as blacklisted IP addresses, remnants of identified malicious files, or URLs classified as phishing sites (Gaucheler 2023). OSINT is a crucial part of producing IoC, providing in-depth information about potential threats.

2. Methodology

The research methodology used included an integrative literature review, critically examining various sources to form a comprehensive understanding of the topic. Throughout the study, several tools designed for OSINT were tested, as recommended in research articles and blog posts. The testing yielded examples of data types generated through intelligence gathering and possibilities for collecting threat information. The tools were tested either in their web interface or a virtual Kali Linux environment. Usage instructions for the tools were generally found in the tool's documentation, and a list of tools used in the study is provided in the bibliography.

The study integrated threat intelligence from OSINT into the research on information sharing, along with relevant legislation and ethics. Sources included legal documents, directive proposals, and scientific articles and research written by industry professionals. For information retrieval, search engines such as Google, Google Scholar, and DuckDuckGo were primarily used. Examples of search terms used in these search engines include "OSINT," "OSINT in threat intelligence," "OSINT regulations," "OSINT ethics," "Threat intelligence sharing," and "Open-Source intelligence legislation." Specifically, in the section focusing on OSINT techniques and benefits, research literature was complemented with research material found in blogs and articles by industry experts, providing a unified perspective. Sources were objectively analyzed using content analysis methodology, and the study aimed to utilize the most recent information available.

3. OSINT for Cybersecurity

OSINT refers to intelligence derived from collecting, evaluating, and analyzing publicly available information. Various entities, including cybersecurity professionals, malicious hackers, and state intelligence services, employ advanced techniques to extract specific information from vast datasets to achieve their objectives (SANS Institute, 2023; Fruhlinger et al., 2023). Intelligence involves the collection of national or international threat information, serving to warn about potential threats or opportunities, gathering information on individuals, or contributing to counterintelligence efforts (Office of the Director of National Intelligence, 2023). OSINT is utilized in several ways (SANS Institute, 2023):

- Security and Intelligence: Gathering information on potential threats to security, such as terrorism or cyberattacks, and collecting intelligence on foreign governments, businesses, or individuals.
- Business and Market Research: Collecting information on competitors, industry trends, and consumer behavior for strategic business decisions.
- Investigative Journalism: Gathering information on politics, business, and criminal activities.
- Academic Research: Researchers use OSINT to collect information on trends, opinions, and economic indicators.

OSINT is employed by governments, law enforcement agencies, defense forces, investigative journalists, human rights researchers, private investigators, law firms, corporate cybersecurity and surveillance teams, and penetration testers. It is also used in private life, such as in people searches (Maor, 2022). OSINT is collected from various sources, including libraries, news, articles, websites, social media, blogs, search engines (including specialized ones like Shodan), the deep web, the dark web, images, videos, source code, or metadata (Kaspersky, 2023; SANS Institute, 2023).

Information can be categorized into structured and unstructured types. Structured sources offer defined data in a specific format, while unstructured sources provide data primarily in text form. OSINT techniques can be passive, semi-passive, or active (SANS Institute, 2023):

- Passive: No interaction with target systems, including scanning public websites or retrieving data from open application programming interfaces.
- Semi-Passive: Involves network scanning, directing traffic to target servers to obtain information while mimicking normal internet traffic.
- Active: Interaction or engagement with the target, such as adding a target as a friend on social media or sending messages.

Organizations can tailor their responses to potential cyber threats by proactively using OSINT to understand threat actors' tactics, techniques, and procedures. Shifting from reactive to proactive defense, organizations can protect themselves from cyberattacks by detecting anomalies through OSINT (Slinde, 2023). Information beneficial to attackers includes open ports, vulnerable and outdated software, poorly defined cloud storage, credentials in source code, and system details like device names, IP addresses, and configurations. External sources like social media and information shared by suppliers and partners can also be relevant for attack planning (Yadav, Kumar & Singh, 2023; Imperva, 2023).

4. OSINT Tool Testing

This study examined tools suitable for OSINT investigations, focusing on commonly used tools, with an emphasis on free and open-source tools. It explored various techniques such as search operators, regular expressions, and multipurpose tools. The OSINT Framework is presented as a comprehensive reference for OSINT tools that provide a wide range of resources for intelligence gathering. The study covers search engine operators and provides examples of commonly used operators to refine search results. It tested the OSINT Framework,

Maltego, TheHarvester, SpiderFoot, ThreatMiner, and other tools for various aspects of OSINT investigations, such as social media analysis, website technology detection, and domain name investigation. Techniques like regular expressions were looked at for structuring and extracting specific data from unstructured data. Multipurpose tools such as Maltego, TheHarvester, and SpiderFoot were tested for their abilities in mapping relationships, gathering information about organizations, and conducting passive intelligence.

The study also tested search engines such as Shodan, TinEye, and Grep.app, which specialize in finding connected devices, reverse image search, and finding code repositories. We tried tools like Have I Been Pwned to check compromised email addresses. Also covered were tools such as Wappalyzer for detecting website technologies, PhishTank for identifying phishing sites, and DNSdumpster for investigating domain names. The discussion ranged from retrieving historical data using the Wayback Machine and Intelligence X to exploring the dark web using Tor web search engines such as Ahmia and Dark.fail. The importance of tracking cryptocurrencies in investigations was also discussed and challenges and opportunities in tracking transactions and addresses in the field of digital currencies were highlighted.

5. Using OSINT Information to Defend Against Cyber Threats

In the examples of the use of the tools, it was noticed that when inquiring about open data sources, information can be found, e.g., about the target organization's employees, finances, IP addresses connected to the organization, the technologies it uses, vulnerabilities, the relationships of persons or organizations with other persons or organizations and their physical locations. By combining this information, a threat actor can plan an effective and efficient attack. The information can also cause reputational harm if it ends up in the wrong hands.

The organization should identify what kind of threats it faces, what could be the motivations of a potential threat actor to carry out a cyber-attack on the organization, and what kind of information it perceives as a threat if it ends up being public. Once the threat area has been identified, the organization can use open data sources to find out what dangerous information is available about them. The organization must have a process by which, after the discovery of potentially dangerous information, follow-up measures are taken to reduce the risk (Hiremath 2023; Rose 2020)

Investigating open data sources is not a one-time project but should be done regularly. Technology is helpful, and the various scans should be automated so that dangerous information is detected before the threat actor and there is time to react before it is too late (Hiremath 2023; Makrushin 2022). Cooperation within the cybersecurity community is important. Sharing OSINT findings and cyber CTI (CTI) with colleagues in the field facilitates joint preparedness against changing threats. Appropriate platforms, such as ISACs (Information Sharing and Analysis Center), are essential here. (Hiremath 2023; ENISA 2023)

6. Sharing CTI in a Trusted Environment

The DYNAMO platform offers support for all phases of cyber resilience. One part of the platform is the ability to store and share CTI among DYNAMO stakeholders, paying special attention to protecting shared information. For this purpose, the project develops a trust environment for users of the DYNAMO platform. The developed module is based on open-source solutions such as the Malware Information Sharing Platform (MISP), taking advantage of its advanced storage and sharing functions, but also allowing integration with other open-source tools such as Cortex and TheHive (DYNAMO 2023). Threat actors specialize in targeting their attacks on certain sectors, such as governments or banking. Once they find a technique that works against one organization, a common next step is to try the same techniques against similar organizations. Therefore, shared CTI in an environment of trust benefits all parties and reduces the probability that similar organizations fall victim to the same threats. (Postolovski 2023)

MISP open-source threat intelligence and sharing platform’s goal is to prevent and identify targeted cyber-attacks, for example with the help of indicators of compromise. MISP gives CTI a unified structure and automatically combines similar information, making it easier to store and share information despite the large amount of information among organizations that are likely to face similar threats. MISP includes a representative state transfer (REST) application programming interface (API) for adding functionalities and external data sources (CIRCL 2023; Postolovski 2023). TheHive is an open-source Security Incident Response platform, i.e. a platform for managing information security incidents. It can be synchronized with MISP instances to investigate MISP events. The research results can be exported and distributed as a MISP event among the trust environments. Cortex is a software developed by TheHive Project for analyzing compromise indicators, which is intended specifically for use with MISP and TheHive.

When sharing CTI, a balance must be found between transparency, privacy, and confidentiality. It is good for the organization to share CTI, which helps other organizations in a trusted environment to understand the threats better. On the other hand, sharing identifiable or sensitive information should be avoided. Such information includes, for example, company names, internal IP addresses, personnel names, customer identifiers, or business-related information. The information to be shared therefore contains intelligence information only about threats, and not about the company's own business, infrastructure, employees, or customers. Checking shared CTI is important to maintain privacy and confidentiality. To reduce risks and ensure confidentiality and accuracy when sharing information, an approval chain can be used before sharing information with other organizations (Postolovski 2023).

7. Conclusions

Open-source intelligence (OSINT) is an important tool in maintaining the cybersecurity of organizations. It can be used to proactively identify and anticipate potential threats and develop measures to reduce these threats. This concerns finding information that is publicly available and that threat actors could use against the organization. This may include, for example, information related to employees, which could be used for example in phishing or password cracking.

Practicing OSINT requires expertise. The professional must be able to identify significant data from a large amount of information and assess its relevance in terms of the organization's security. Ethical and legislative aspects are central to the collection of intelligence information. The data collection must be limited so that it does not violate anyone's privacy or rights. The use of technology in data collection can lead to unethical or illegal practices if people's control and understanding of the limitations of intelligence are not a priority.

The inquiry process must have a clear reference framework, in which the goals and processes of OSINT are defined. The role of technology in data collection is significant in speeding up data acquisition, but the importance of human judgment is emphasized in ethical decisions. Documenting operations is central to ensuring ethics. Automated systems can save research steps, allowing people to focus on analysis and value creation.

Platforms such as MISP are used to share threat information, the goal of which is to prevent and identify targeted cyber-attacks and share CTI between actors. When sharing CTI, it is important to ensure that the shared information does not violate privacy or confidentiality. Operators related to data security can share information within the framework of the Personal Data Act when the activity serves the public interest and personal data is only processed to the extent necessary. In addition, the requirements set by, for example, the revision of the Network and Information Security Directive (NIS2) directive and the artificial intelligence regulation must be considered as they impose several new cyber security obligations on organizations in many economically critical sectors.

OSINT offers organizations an important tool for maintaining cybersecurity, but its effective and ethical use requires expertise, clear processes, and a balanced combination of technology and human skills. The ethics of OSINT activities and the utilization of artificial intelligence are topics that would be enough to consider for separate further studies. These two issues could also be combined into one research entity.

Acknowledgements

Acknowledgement is paid to DYNAMO Project, funded by the European Union under grant agreement no. 101069601. Views and opinions expressed are however 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.

References

CIRCL. 2023. CIRCL » MISP - Open Source Threat Intelligence Platform. Referred 24.10.2023. https://www.circl.lu/services/misp-malware-information-sharing-platform/

DYNAMO. 2023. Horizon Dynamo EU. Referred 21.9.2023. https://horizon-dynamo.eu/about/

ENISA. 2023. Information Sharing and Analysis Centers (ISACs). Topic. ENISA. Referred 23.10.2023. https://www.enisa.europa.eu/topics/national-cyber-security- strategies/information-sharing

Fruhlinger, Josh, Sharma, Ax & Breeden, John. 2023. 15 top open-source intelligence tools. CSO Online. Referred 11.9.2023. https://www.csoonline.com/article/567859/what-is-osint- top-open-source-intelligence-tools.html

Gaucheler, Mathieu. 2023. Advanced IOCs Collection with OSINT and Threat Intelligence Feeds. Maltego. Referred 11.9.2023. https://www.maltego.com/blog/advanced-iocs- collection-with-osint-and-threat-intelligence-feeds/

Hiremath, Omkar. 2023. Protecting Your Organization With Open-source Intelligence (OSINT). Software Secured. Referred 23.10.2023. https://www.softwaresecured.com/protecting-your- organization-with-open-source-intelligence-osint/

Imperva. 2023. Open-Source Intelligence (OSINT). Referred 5.9.2023. https://www.imperva.com/learn/application-security/open-source-intelligence-osint/

Kaspersky. 2023. OSINT (Open-Source Intelligence). Referred 5.9.2023. https://encyclopedia.kaspersky.com/glossary/osint/

Kumar, Dinesh. 2022. 10 OSINT Tools Hackers Need to Know About. Medium. Referred 1.10.2023. https://0xtmux.medium.com/10-osint-tools-hackers-need-to-know-about- 9cbb3519ea47

Maor, Etay. 2022. How and Why to Apply OSINT to Protect the Enterprise. Dark Reading. Referred 23.10.2023. https://www.darkreading.com/attacks-breaches/how-and-why-to-apply-osint-to-protect-the-enterprise

Office of the Director of National Intelligence. 2023. What is Intelligence? Referred 21.9.2023. https://www.dni.gov/index.php/what-we-do/what-is-intelligence

Packham, Karen. 2022. RHEA to Contribute to Develop Cyber Threat Intelligence Solution for European DYNAMO Project. RHEA Group. Referred 11.9.2023. https://www.rheagroup.com/rhea-to-contribute-to-develop-cyber-threat-intelligence- solution-for-european-dynamo-project/

Postolovski, Tash. 2023. What is MISP? The Ultimate Introduction. Cosive. Referred 26.10.2023. https://www.cosive.com/blog/what-is-misp-the-ultimate-introduction

SANS Institute. 2023. What is Open-Source Intelligence? Referred 28.8.2023. https://www.sans.org/blog/what-is-open-source-intelligence/

Slinde, Johanna Sofie. 2023. Unveiling the Potential of Open-Source Intelligence (OSINT) for Enhanced Cybersecurity Posture

Rose, Zoë. 2020. OSINT – Using Threat Intelligence to Secure Your Organisation. Tripwire. Referred 11.9.2023. https://www.tripwire.com/state-of-security/osint-using-threat-intelligence-secure-organisation

Implementation of OSINT for Improving an International Finance Sector Organization's Cybersecurity

Jyri Rajamäki and Krista Tiitta
Laurea University of Applied Sciences, Espoo, Finland

jyri.rajamaki@laurea.fi
krista.tiitta@student.laurea.fi

Abstract: This work-in-progress paper addresses the need to improve intelligence processes and enhance an organization's response to cyber threats while managing associated risks and improving Business Continuity Management (BCM). The paper focuses on the role of Open-Source Intelligence (OSINT) in Cyber Threat Intelligence (CTI) gathering and presents an operational process for its implementation. The process includes defining goals, selecting open sources, data collection, filtering, analysis, and reporting. Testing in an international financial sector organization yielded positive results, demonstrating the process's value in threat intelligence. Future research should clarify the role of artificial intelligence in OSINT.

Keywords: Business continuity management, Cybersecurity, Cyber threat intelligence, Design science research, OSINT

1. Introduction

The fundamental pillars of information and cybersecurity include ensuring and protecting the confidentiality, integrity, and availability of data. Organizations are tasked with securing their information as effectively as possible, which also includes protecting their IT infrastructure and personnel while minimizing the attack surface. Failure to secure and protect information can have significant financial implications, additional work for staff, legal repercussions, and damage to the organization's reputation (Gatlin, Yampolskiy & Yung, 2021). Cyber Threat Intelligence (CTI) is information based on knowledge, skills, and experience that helps reduce potential attacks and malicious events in cyberspace (DYNAMO, 2023).

The case organization - an international player in the financial sector - had recognized weaknesses and deficiencies in utilizing open-source intelligence (OSINT), and there were no established processes related to it. For this reason, a Design Science Research (DSR) project shown in Figure 1 was launched to design a new operational process for the case organization.

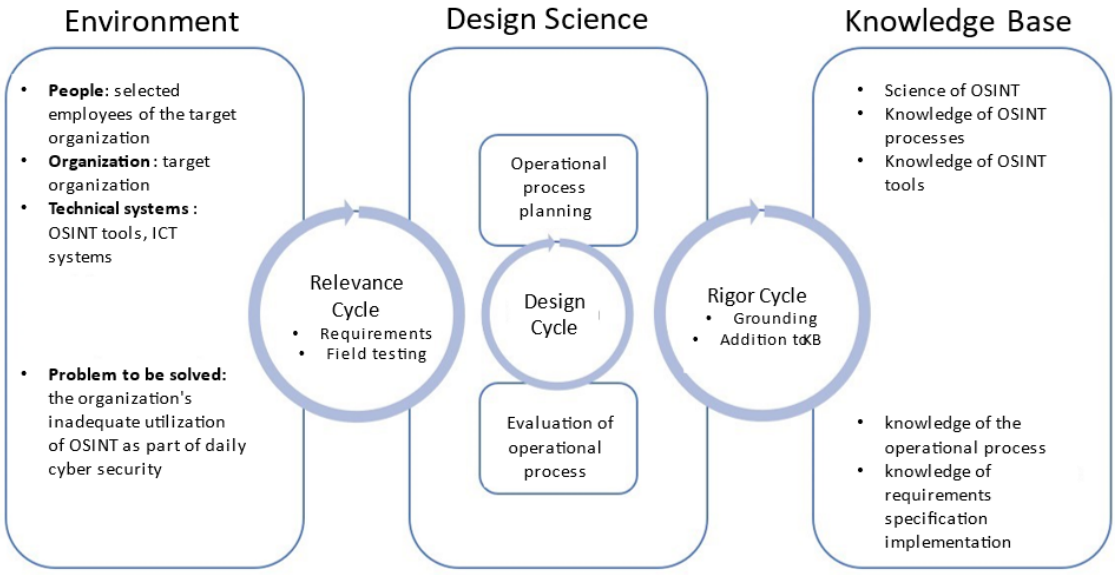


Figure 1: The DSR framework applied in the study (modified from Hevner, 2007)

To maintain anonymity in this study, the requirement specification for the OSINT intelligence operational process is presented in a simplified manner. The requirement specification defines the needs and starting point for process planning. The starting point and needs for process planning in this study are based on the case organization's requirements and the real problem. The goal is to develop a functional, clear, and efficient operational OSINT process. The requirement specification also includes a general description of the process,

which covers the scope of the operational process, functionalities, users, constraints, functional requirements (what the process includes and what is required for it to meet its intended purpose), and non-functional requirements (scalability, performance, and usability, i.e., qualitative requirements).

2. OSINT

Open Source Intelligence (OSINT) is a multifaceted concept that encompasses the systematic collection and analysis of information obtained from publicly accessible and legally obtained sources. These sources span a broad spectrum, ranging from the digital realm, including the Internet and social media, to more traditional media channels such as newspapers, magazines, television, and radio. The key characteristic of OSINT is its inclusivity, as it extends to any information or data that is available to the public. This inclusivity is particularly relevant in the contemporary era, given the exponential growth of open-source information facilitated by the widespread use of digital platforms, with social media platforms being a prominent contributor to this abundance of data (Akhgar & Bayerl, 2015; Azzopardi et al., 2014).

The pervasive use and accessibility of the Internet have fundamentally altered the way individuals and organizations communicate and share information. The advent of social media platforms and real-time messaging services has ushered in a new era of instantaneous interaction. However, this convenience comes with risks, as it has led to the inadvertent exposure of sensitive information. This unintentional disclosure has raised concerns about the potential misuse of such data, including its exploitation for nefarious purposes like harassment, terrorism, or identity theft (Azzopardi et al., 2014).

OSINT has evolved into a pivotal tool in various security and intelligence operations. It finds applications in diverse fields, including criminal investigations, counterterrorism efforts, the monitoring of Advanced Persistent Threat (APT) actors (i.e., sophisticated hacking groups that may be state-sponsored), and the fight against cybercrime. OSINT's growing significance can be attributed to its cost-effectiveness and reduced risks when compared to traditional espionage methods (Oikarinen, 2020).

Within the realm of information security, OSINT intelligence serves as a valuable resource for identifying vulnerabilities within an organization's IT infrastructure. The concept of a "Red Team" exemplifies this approach, wherein a group simulates adversarial attacks against an organization. These simulated attacks include an initial phase of information gathering, a process that often relies on OSINT techniques (BreachLock, 2023).

Open Source Intelligence plays a crucial role in penetration testing and Red Teaming exercises. It aids in the identification and exposure of potential vulnerabilities and resources that may have inadvertently leaked outside an organization's security perimeter. These findings empower security experts to proactively address vulnerabilities and mitigate associated risks before malicious actors can exploit them. Some common vulnerabilities identified through OSINT intelligence include the inadvertent exposure of sensitive data, the use of outdated and unpatched software, and the existence of open ports on IT devices (BreachLock, 2023).

OSINT is often characterized as providing an "attacker perspective" because it offers cybercriminals insights into the same vulnerabilities and findings that security professionals may discover. Once attackers identify a vulnerability, exploiting it becomes relatively straightforward. Consequently, many small and medium-sized organizations are susceptible to cyberattacks, with attackers often combining OSINT intelligence with social engineering techniques in targeted phishing campaigns (BreachLock, 2023).

Threat Intelligence is a critical component in an organization's cybersecurity strategy. It is the practice of proactively identifying, assessing, and mitigating threats targeting the organization. Effective threat intelligence relies on the systematic collection and analysis of data, including publicly available information relevant to the organization. The threat intelligence process typically involves six stages: defining objectives, data collection, data processing, data analysis, reviewing results, and providing feedback (Martins & Medeiros, 2022).

In today's interconnected digital landscape, OSINT serves as a valuable resource for security professionals seeking to gain a comprehensive understanding of potential cyber threats. The vast expanse of the internet offers a wealth of information on a wide range of topics, including current threats, emerging trends in the cyber domain, and the evolving tactics and techniques employed by cybercriminals. Information security experts and their teams can leverage data extracted from open sources to protect their organization's IT infrastructure effectively. This includes preemptively addressing zero-day vulnerabilities and implementing additional security measures to safeguard against potential threats (BreachLock, 2023).

In summary, OSINT is a versatile and invaluable tool within the domain of cybersecurity. It facilitates the systematic gathering, analysis, and utilization of information from publicly accessible sources, enabling organizations to enhance their threat assessment and mitigation strategies.

3. OSINT Operational Process

In the case organization, the OSINT operational process will primarily be used for threat intelligence purposes. Ideally, threat intelligence is a part of every organization's cybersecurity strategy. Monitoring existing and potential new threats, consistent data collection, and analysis enhance an organization's ability to protect itself from cybercrime. The Threat Intelligence Lifecycle consists of six phases: defining objectives for threat intelligence, data collection, data processing, data analysis, results review, and feedback (Snyk 2023). Figure 2 presents the operational process designed for the case organization.

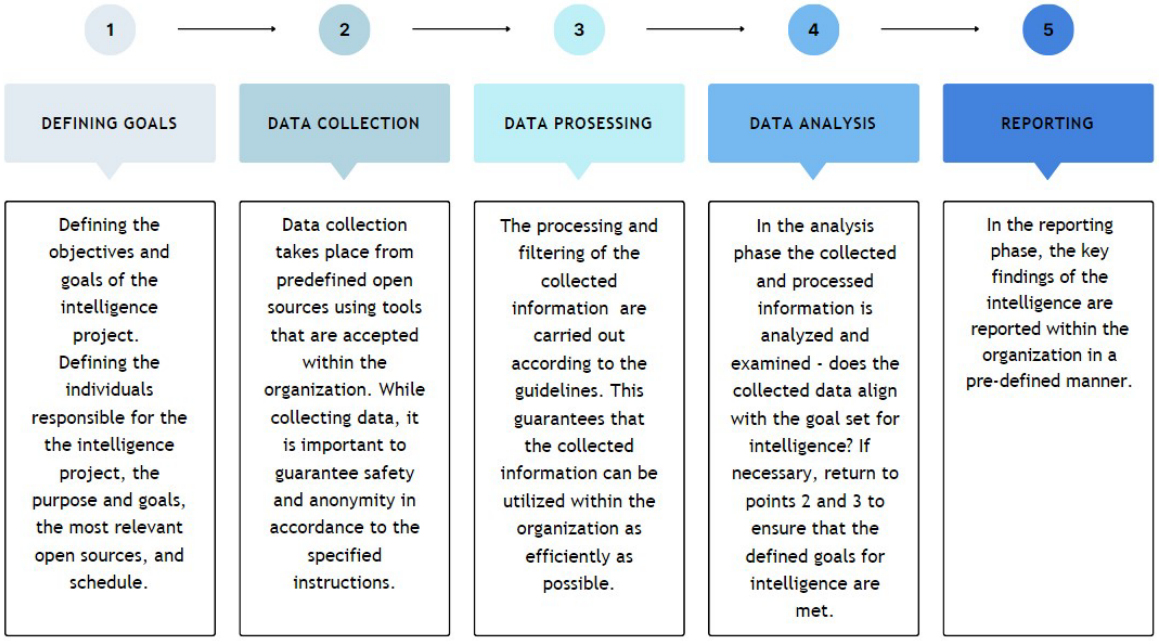


Figure 2: Designed OSINT Operational Process

The operational process begins with defining the goals of intelligence gathering. In this phase, the purpose of intelligence gathering is determined, and the individuals responsible for gathering the intelligence are selected. Intelligence gathering is mainly carried out by pre-designated employees within the organization who have an intelligence orientation. Additionally, the most relevant open sources are selected to maximize the effectiveness of intelligence gathering. At the same time, the intelligence project's timeline is defined.

During the data collection phase of the process, information is gathered from the open sources identified in the first phase of the operational process. Data is collected using approved tools within the organization for intelligence purposes. Common tools used for OSINT intelligence include Maltego and SpiderFoot. When conducting active intelligence gathering in addition to passive intelligence gathering or when collecting data from the dark web, one's own actions must be protected as defined within the organization.

In the third phase of the process, during the data collection phase, the collected data is filtered and processed. Filtering and processing of collected data are carried out according to the organization's guidelines. The collected data should align with the intelligence objectives set for the process.

In the fourth phase of the process, the processed and filtered data is analyzed and examined. At this stage, it is determined whether the collected data meets the intelligence objectives. If the analyzed data aligns with the intelligence objectives, the process can proceed to the final phase (reporting). However, if it is determined that the analyzed data does not meet the intelligence objectives, steps two and three of the process are repeated to achieve the intelligence objectives, ensuring that the analyzed data is accurate and its utilization is efficient. If steps two and three need to be repeated, new open sources may also be identified from which information is sought.

The final phase of the operational process is the reporting phase. At this stage, the analyzed information is reported. The findings of the intelligence gathering are reported in a manner defined within the organization to pre-determined individuals and/or units. The produced report should be clear and concise so that even individuals who are not technology-oriented can understand the information presented in the report. Based on the report and the information gathered in the intelligence process, potential follow-up measures are defined to enhance the organization's security.

4. Testing of the Operational Process

The operational process was tested by conducting an OSINT investigation on an organization's employees. Permission was obtained from the individual before conducting the investigation. Clear objectives were defined for the intelligence gathering. The goal was to gather as much information as possible about the individual, which could be used for targeted phishing attacks or other forms of manipulation. This paper's main author conducted the OSINT investigation. The timeline was set for April 2023, but the intelligence objectives were achieved in early April. Social media primarily served as the source of information. Specific tools were designated for this OSINT investigation, including Maltego, SpiderFoot, Google, and Google Maps. Data collection was swift, and the information that aligned with the intelligence objectives was found relatively effortlessly. The collected data was analyzed to ensure its relevance. The relevance of the information was cross-verified through multiple sources.

There was no need to revisit steps two and three of the operational process in this investigation. Sufficient information that matched the intelligence objectives was collected from the individual. The quality of the information was such that it could have been used for actions like targeted phishing or other forms of manipulation. Following the analysis of the information, a report was created from the findings of the intelligence and presented to the subject of the investigation. Subsequently, the intelligence report was reviewed with the target individual, who was provided with guidance on improving their online security. Corrective actions in the individual's online activities did not require major changes.

5. Discussion

The objective of this study was to implement OSINT as a component of the case organization's cybersecurity. The implementation was achieved by designing an operational process for OSINT intelligence, enabling the organization to enhance its cybersecurity management. The research methodology was design science. The design of the operational process closely adhered to the seven principles of DSR (Hevner et al, 2004), and the research fulfilled the requirements outlined in these principles as follows:

- An operational process for OSINT intelligence was created for the case organization to address their need to improve the deficient and weak approach to utilizing open-source intelligence within the organization.
- The resulting operational process solves the case organization's problem related to the inadequate utilization of OSINT.
- The operational process was thoroughly tested, and it was found to meet the established objectives for its functionality.
- The operational process provides the case organization with a customized process for conducting OSINT intelligence within the organization. The process is based on scientific knowledge.
- A comprehensive study was conducted on the scientific theory related to OSINT intelligence, requirement specification theory, and process design theory.
- The operational process was created by using scientifically valid knowledge and applying an existing solution to the intelligence process.
- The results of the study were presented within the case organization to individuals for whom the research findings were relevant. The research results were also shared within the academic community.

The study results were validated through testing the developed operational process, and the outcomes aligned with the objectives set for intelligence gathering and testing. The case organization can utilize this operational process, particularly for threat intelligence purposes. Therefore, it can be concluded that the study successfully generated value for the case organization, achieving its research objective, and the design science research approach was effective.

As future research, it would be interesting to investigate how different financial industry players utilize OSINT intelligence and its processes to maintain a threat landscape, acquire targeted threat intelligence, and mitigate the initial stages of attack vectors using OSINT. The focus of such research could be on intelligence gathered from the dark web, specifically what information about financial industry entities can be found there and how this information can be strategically used by cybercriminals against financial institutions. This could be examined through frameworks like Lockheed Martin's Cyber Kill Chain, which is part of the Intelligence-Driven Defense model used to identify and prevent cybercriminal activities (Lockheed Martin 2023).

Another area for further research could be the challenges and opportunities brought by artificial intelligence (AI) in OSINT intelligence. What benefits does AI offer to OSINT intelligence, and what challenges may arise as AI technologies continue to evolve?

Acknowledgements

This work was supported by the DYNAMO project, which has received funding from the European Union's Horizon Europe research and innovation funding programme under grant agreement no. 101069601. The sole responsibility for the content of this paper lies with the authors. It does not necessarily reflect the opinion of the European Commission or of the full project. The European Commission is not responsible for any use that may be made of the information contained therein.

References

Akhgar, B. & Bayerl, P. (2015) "Surveillance and Falsification Implications for Open Source Intelligence Investigations", *Communications of the ACM*, Vol 58, Iss 8, 62.

Azzopardi, L., Glisson, W., Maxwell, D., & McKeown, S. (2014) "Investigating people: a qualitative analysis of the search behaviours of open-source intelligence analysts", *Proceedings of the 5th Information Interaction in Context Symposium*, pp 175-176.

BreachLock (2023) "What is Open-source Intelligence, and how is it used?" [online] https://www.breachlock.com/resources/blog/what-is-open-source-intelligence-and-how-is-it-used/

Gatlin, J., Yampolskiy, M. & Yung, M. (2021) "Myths and Misconceptions in Additive Manufacturing Security: Deficiencies of the CIA Triad", *Proceedings of the 2021 Workshop on Additive Manufacturing Security*.

DYNAMO (2023) "Factsheet 2: Cyber Threat Intelligence" [online] https://horizon-dynamo.eu/wp-content/uploads/2023/09/DYNAMO-Factsheet-2-1.pdf

Hevner, A. (2007) "A Three Cycle View of Design Science Research", *Scandinavian Journal of Information Systems* vol 19, pp 88-91.

Hevner, A., March, S., Park, J., & Ram, S. (2004) "Design Science in Information Systems Research", *MIS Quarterly* Vol 28 No 1, pp 80–90.

Lockheed Martin (2023) "Putting Intelligence to Work" [online] https://www.lockheedmartin.com/en-us/capabilities/cyber/intelligence-driven-defense.html

Martins, C. & Medeiros, I. (2022) "Generating Quality Threat Intelligence Leveraging OSINT and a Cyber Threat Unified Taxonomy", *ACM Transactions on Privacy and Security*, Vol 25, Iss 3, Article 19, pp 2-5

Oikarinen, A. (2020) "Open-Source Intelligence – It's Incredible what you can find from public sources" [online] https://www.nixu.com/blog/open-source-intelligence-its-incredible-what-you-can-find-public-sources

Snyk (2023) "Threat Intelligence Lifecycle - Phases & Best Practices Explained" [online] https://snyk.io/learn/threat-intelligence/threat-intelligence-lifecycle/

Art Crime Does not pay: Multiplexed Social Network Analysis in Cultural Heritage Trafficking Forensics

Jarno Salonen[1] and Alessandro Guarino[2]
[1]VTT Technical Research Centre of Finland, Tampere, Finland
[2]StAG SASU, Antibes, France

jarno.salonen@vtt.fi
a.guarino@stagcyber.eu

Abstract: Nowadays, crimes connected to cultural heritage can feature as a staple for organised crime networks and act as financial enablers for international conflicts, including terrorism organisations and even inter-state conflicts, in several ways. Goods of cultural significance include a range of valuable objects related to human cultures, like works of art, historical artefacts, and other antiques, but also forgeries based on such objects. These crimes are almost always transnational, for instance, involving theft or looting in one country and goods moved across borders to be sold. This article presents an intelligence methodology based on Social Network Analysis (SNA) techniques that can support law enforcement agencies (LEAs) in their daily struggle against criminals that also pose a threat to national security. The methodology proposed is based on the building of a blended, multiplexed social network graph, deriving from the fusion of a diverse set of data sources, both in the open-source domain (OSINT) and in the classified domain. We will present data collection methods, correlation between sources, possible ways to generate blended links between individuals that retain information from different sources, and SNA techniques applied to intelligence and investigations. The article provides an answer to the following research questions: how we can detect and identify criminal activities and networks related to cultural goods crimes, how we can assist LEAs in countering illicit trafficking, and how we can ensure that art crime does not pay.

Keywords: Cultural goods crime, Cybercrime, Open-source intelligence, Cybersecurity, Investigation

1. Introduction

There has been a considerable increase in organised looting, trafficking, and the illicit sale of cultural objects during the last decade (Interpol, 2021). The online art market reached 67,8 billion USD in 2022, with almost 38 million art transactions worldwide, while online sales accounted for roughly 16% of the total market value (Statista 2023). A Financial Action Task Force report also confirms that the cultural goods market attracts organised crime groups to fund their activities (FATF, 2023). These objects are expensive and difficult to trace in an opaque market that has moved mostly online. Given their complex nature, contrasting these networks is not easy. The online market enables better opportunities for the criminals through faster online sales, anonymity or falsified seller information, and even bank accounts not connected to the criminals. This makes it impossible for the authorities to detect illegal sales activities, stop monetary transactions, and seize the goods. Therefore, art crimes have an impact on national security since they can often be enablers for threats from terrorist groups, nation-state influence groups, or other forms of trafficking (weapons, drugs or human beings). We present a methodology to even the odds of authorities detecting the illicit trafficking of cultural goods in online marketplaces through approaches based on social network analysis (SNA) (Burcher, 2017). The technology is already in use, but our concept focuses more on the process perspective.

2. A Typical Cultural Heritage Trafficking Scenario

The looting and illicit trade of cultural items is a complex form of crime, transnational in nature, and involves a very diverse set of "stakeholders" (Interpol, 2023). Our scenario includes: i) looting groups close to heritage sites, often in developing countries; ii) figures taking advantage of or at some level abetting the looters, e.g., corrupt police officers, local authorities, heritage professionals; iii) transit professionals: shipping companies, border and customs agents, brokers; iv) art market professionals: legitimate actors like auction houses, dealers, collectors and museum employees; v) facilitators: for instance, restorers, authentication laboratories (who certify the provenance of items), and academics. These stakeholders are connected in a loose network of personal relations spanning over several countries and crossing the line between legal and illegal, licit and illicit activities. Looters in the origin country operate based on their local knowledge as well as on prompt advice from their supporting contacts; for instance, they might remove an ancient statue from an archaeological site. Supporting contacts are also a necessary part of the trade flow, collaborating, e.g., in moving the looted statue out of the country and into the destination market without interference from LEAs and authorities who, being unaware of the existence of the looted object, cannot investigate any crime in the original country before the statue is exported. The shipping phase might see the involvement of corrupt customs officers in the exporting countries and possibly in the transit and destination countries as well. Once it reaches the destination country, the statue could enter the

legitimate antiquities market with forged provenance and be put on sale by a legitimate auction house, where it is eventually acquired by a private collector, usually kept anonymous. Investigations of such cases involve crimes committed both in the original and destination countries and are often triggered by the claims of the original country government, which recognises the object as looted once it appears on the auction or gallery catalogue.

3. Background and Data Sources

In this section, we describe the possible data sources that can feed the blended social network graph (SNG). SANS (2023) defines open-source intelligence (OSINT) as "intelligence produced by collecting, evaluating, and analysing publicly available information with the purpose of answering a specific intelligence question". One of the most relevant OSINT techniques in our ongoing project is harvesting data from online sources such as websites, social media, and other online repositories/databases. They represent the most important sources for the creation of the SNG. Additional sources, such as call data records (CDRs) and geospatial intelligence (GEOINT) information, are also used to detect criminal activity, especially at cultural heritage sites. There is previous research on criminal investigation using CDR data (Khan et al., 2017) as well as on the use of CDR data in large-scale human and urban mobility (Thuillier et al., 2018). Data on cryptocurrency and other financial transactions can also be useful in criminal investigations, which has been studied among others by Nikkel (2020) and Kethineni and Cao (2020).

LEAs use many types of public records even outside of the ongoing criminal investigations, including court sentences, company registration records, tax records, etc. The challenge with these data sources is processing unstructured data into structured formats suitable for the building of an SNG layer. CDRs can be lawfully obtained by LEAs in most countries. CDRs contain sensitive information like phone numbers, terminal and account identifiers (IMEI, ICCID), and possibly IP addresses, and they may also provide link information and their significance (number of calls, texts between accounts, etc.). Existing digital evidence is extremely useful. For example, forensic images of seized phones or laptops yield treasure troves of information about the users and their contacts, providing ready-made correlations between call records, social media contacts, email messages, and more. Such sources, however, come with legal and ethical limitations to their use, even in the context of an existing criminal case.

4. Building the Multiplexed Social Network Graph

The SNG is the core data structure of the methodology. Nodes represent individuals (e.g., members of the looting criminal group, smugglers, art dealers) and their features (e.g., name, phone number) in the graphs. Links between individuals represent the strength of their relationship. In a SNG built using a corpus of email messages, the weights of the links are represented by the number of messages exchanged, and they can be either directional or not. Links are directional when it is important to consider the direction of the dependence between a node or individual and the others. For example, in the case of the members of a social network, a directional link represents the "follower to followed" relation. Our methodology aims at building a multiplexed SNG, which is the blend of the SNGs resulting from the processing of each data source. Of course, the graphs will not overlap precisely, so the challenge is to correlate the information pertaining to the same node (individual) across different data sources. To achieve this, AI-based techniques, manual input from expert operators, or a combination of the two can be used. Conflicts among sources should be solved, taking into consideration the reliability of each source that analysts evaluate. The result of the merging is a list of features associated with a single node. The individuated nodes could very well not present links from the entirety of the data sources, but that should not represent a problem. The actual challenge is to identify the relative weights to be assigned to each feature as well as their reliability. The intelligence tool implementing this methodology will be able to remove the SNG "layers" from one or more sources at any moment in time to accommodate the needs of the investigator or intelligence analyst. In our scenario, information about the individual looters may also come from CDRs (phone number, calls) and be correlated with social media posts or previous cases investigated by the LEAs.

5. Analysis of the Social Network Graph

A timely updated, content-rich SNG, timely updated with new information, is the object on which analysis tools can be brought to bear. We need to make a distinction between the use of our methodology by LEAs in the context of criminal investigations and its use as an intelligence tool by security agencies. In the first case, the SNA methodology is intended to support the intelligence-led policing paradigm explained in Burcher (2020). However, in this context, limitations in the collection and use of data imposed by privacy and data protection

laws, among others, could impact the insights that can be gained. Moreover, in this use case, the results of the analysis should be traceable and auditable to ensure a proper evidence chain of custody and the integrity of the potential digital evidence. In the second case, such limitations could, in many cases, be irrelevant.

Well-established measures and algorithms can be fruitfully employed to extract meaningful intelligence and investigative leads to efficiently target specific nodes. Of course, the first useful tool is the actual graphical visualisation of the network. Visual analysis is surprisingly helpful in the hands of a skilled operator. Centrality measures identify relevant nodes, albeit in different ways. The simplest kind (degree centrality) counts the number of links (degree) of a node. The most connected nodes could be, for instance, the leader of one organisation or part of it. In our scenario, for instance, the local leader of the looting group. Betweenness centrality measures an interesting aspect of complex networks as it has the power to identify "brokers", i.e., individuals that are not connected to many other nodes but lie between two distinct sections of the network that are internally highly connected. Mathematically, betweenness gauges (roughly defined) how nodes can reach out to the entirety of the network in an efficient way. The individuals identified by this measure could be, for instance, the "cut-off" between the criminal part of the whole network (looters, smugglers) and the legal one (art market, collectors), or the persons tasked with delivering messages and instructions from the leadership. In other words, intermediaries of information or power. Targeting these persons of interest could disrupt the whole network with a minimal use of resources. More complex algorithms can identify a consistent part of the whole network, for instance, because they are highly connected internally but weakly connected to the rest of the network. These are only a few examples of the potential analysis tools that, when employed on a nuanced, multi-layer SNG, can support investigations.

6. Discussion

The presented SNG methodology and especially the correlation engine, is currently still under development, and the final version will be ready early next year, after which it will be tested and evaluated by end-users. We have already identified a few challenges, one of which is the development of efficient AI-enabled algorithms for correlation definition. Another challenge is the lawful collection of massive amounts of data, which we are trying to overcome by producing artificial test data sets, which we aim to publish for future research and development needs.

7. Conclusion

We present ongoing research on SNA-based methodology to support criminal investigations on art crimes. The use of our methodology by LEAs will enable them to use their limited resources in the most efficient way and disrupt the criminal networks responsible for these kinds of crimes. The objective is to use OSINT techniques for harvesting data from online sources such as websites, social media, and other online repositories and using them to create a blended social network graph (SNG). Additional sources, such as call data records (CDR) and geospatial intelligence, can be used to enrich the graph, as well as data on cryptocurrency and other financial transactions. Nodes within the SNG represent individuals, such as the members of the criminal group; node features contain names, phone numbers and other relevant information; and the links and their weight represent the strength of the relationships between the nodes. With the use of our methodology, LEAs can enhance their investigations, discover criminal networks at a faster pace, and counter the illicit trafficking of cultural heritage goods, thus ensuring that art crime does not pay.

Acknowledgements

This research has been fostered by the RITHMS project (G.A. 101073932), funded by the European Union. The views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

References

Burcher, M., Whelan, C. (2017) "Social network analysis as a tool for criminal intelligence: understanding its potential from the perspectives of intelligence analysts" - Trends in Organised Crime

Burcher, M. (2020) "Social Network Analysis and Law Enforcement - Applications for Intelligence Analysis" - Palgrave Macmillan

FATF (2023), "Money Laundering and Terrorist Financing in the Art and Antiquities Market" [online], FATF, Paris, France, https://www.fatf-gafi.org/publications/Methodsandtrends/Money-Laundering-Terrorist-Financing-Art-Antiquities-Market.html

Interpol (2021) "Assessing crimes against cultural property - survey of interpol member countries", [online], https://www.interpol.int/content/download/16751/file/2020%20Assessing%20Crimes%

Interpol (2023) "The issues – cultural property", [online], https://www.interpol.int/Crimes/Cultural-heritage-crime/

Kethineni, S. and Cao, Y. (2020), "The rise in popularity of cryptocurrency and associated criminal activity", International Criminal Justice Review, 30(3), 325-344.

Khan, S., Ansari, F., Dhalvelkar, H. A. and Computer, S. (2017). "Criminal investigation using call data records (CDR) through big data technology", in 2017 International Conference on Nascent Technologies in Engineering (ICNTE) (pp. 1-5). IEEE.

Nikkel, B. (2020). "Fintech forensics: Criminal investigation and digital evidence in financial technologies", Forensic Science International: Digital Investigation, 33, 200908.

ANS. (2023). "What is Open-Source Intelligence" [online]. SANS blog, February 23, 2023, https://www.sans.org/blog/what-is-open-source-intelligence/

Statista. (2023). "Art market worldwide - statistics facts", [online]. S. R. Department. https://www.statista.com/topics/1119/art-market/

Thuillier, E., Moalic, L., Lamrous, S. and Caminada, A. (2017). "Clustering weekly patterns of human mobility through mobile phone data", IEEE Transactions on Mobile Computing, 17(4), 817-830.